"Recomendo. Um livro essencial." — BILL GATES

NICK BOSTROM

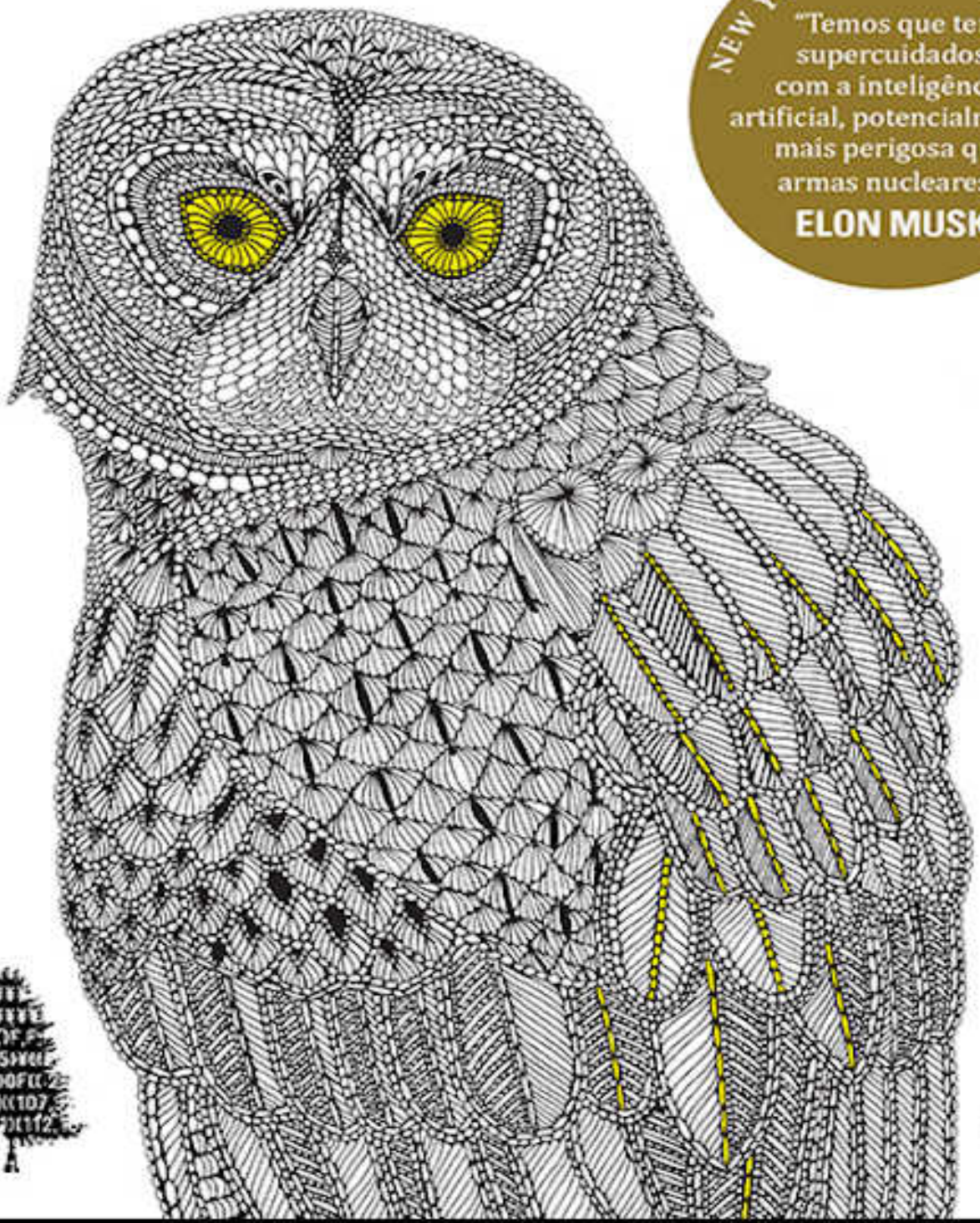
SUPERINTELIGÊNCIA

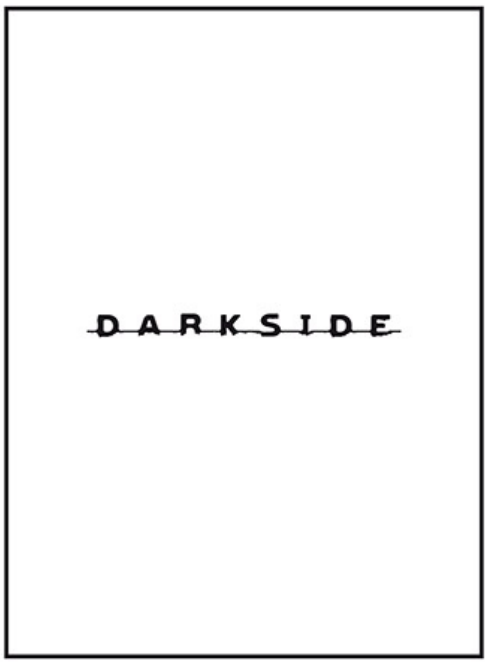
CAMINHOS, PERIGOS E ESTRATÉGIAS PARA UM NOVO MUNDO

NEW YORK TIMES BESTSELLER

"Temos que ter supercuidados com a inteligência artificial, potencialmente mais perigosa que armas nucleares."

ELON MUSK




DARKSIDE





#DARKSIDEBOOKS

SUPER INTELIGÊNCIA

NICK BOSTROM



NICK BOSTROM

Diretor do Instituto para o Futuro da Humanidade
Diretor do Strategic AI Research Centre
Professor da Faculdade de Filosofia & Oxford Martin School
Universidade de Oxford

SUPER INTELI-GÊNCIA

CAMINHOS, PERIGOS E ESTRATÉGIAS PARA UM NOVO MUNDO

Tradução
Aurélio Antônio Monteiro,
Clemente Gentil Penna,
Fabiana Geremias
Monteiro,
Patrícia Ramos Geremias

Revisão Técnica
Bianca Zadrozny, Ph.D.

A FÁBULA INACABADA DOS PARDAIS

Era a estação de construção dos ninhos. Mas, após longos dias de trabalho árduo, os pardais se sentaram sob o luar, relaxando e gorjeando.

“Nós somos tão pequenos e fraquinhos… Imagine como a vida seria fácil se tivéssemos uma coruja que nos ajudasse na construção de nossos ninhos!”

“Sim”, disse outro. “E nós poderíamos ter a ajuda dela para cuidar dos pardais mais velhos e também dos mais novos.”

“Ela poderia nos aconselhar e ficar de olho no gato do vizinho”, acrescentou um terceiro.

Então, Pastus, o pássaro mais velho, disse: “Vamos enviar olheiros para procurar em todos os lugares uma corujinha abandonada, ou talvez um ovo. Um filhote de corvo ou de doninha também serviria. Isso poderia ser a melhor coisa que já nos aconteceu, pelo menos desde a abertura do Pavilhão do Grão Ilimitado, no quintal ao lado”.

O bando estava radiante e os pardais começaram a gorjear com toda a força.

Apenas Scronkfinkle, o pardal rabugento e de um olho só, não estava convencido da prudência daquela empreitada. Disse ele: “Será, com certeza, nossa destruição. Não deveríamos pensar um pouco sobre a arte da domesticação e do adestramento das corujas antes de trazer criaturas desse tipo para o nosso meio?”.

Pastus respondeu: “Adestrar uma coruja parece algo extremamente difícil. Vamos começar procurando um ovo, o que já não será tarefa fácil. Depois que tivermos sido bem-sucedidos em criar uma coruja, então poderemos pensar em outros desafios”.

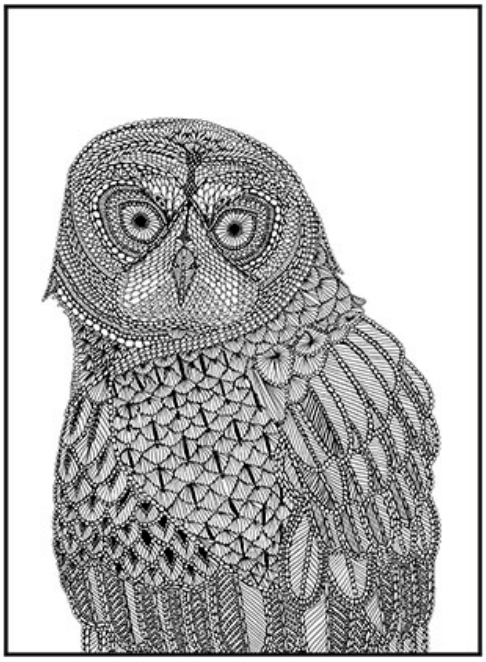
“Há uma falha nesse plano”, grunhiu Scronkfinkle, mas seu protesto foi em vão, pois o bando já havia levantado voo para começar a colocar em ação as diretrizes fornecidas por Pastus.

Apenas dois ou três pardais ficaram para trás. Juntos começaram a imaginar como as corujas poderiam ser domesticadas e adestradas. Logo perceberam que Pastus estava certo: esse seria um desafio extremamente difícil, principalmente porque não havia uma coruja com a qual praticar. Entretanto, eles se empenharam ao máximo, temendo que o bando pudesse retornar com um ovo de coruja antes que uma solução para o problema de contenção fosse encontrada.

O fim da história não é conhecido, mas o autor dedica o livro a Scronkfinkle e seus seguidores.

NICK BOSTROM

SU-
PER-
IN-
TELI-
GÊN-
CIA



SUPERINTELIGÊNCIA

SUMÁRIO

Capa

Mídias sociais

Folha de rosto

Figuras, tabelas e quadros

Prefácio

Capítulo 1

- Desenvolvimentos anteriores e capacidades atuais
  - Regimes de crescimento e a história
  - Grandes expectativas
  - Períodos de esperança e aflição
  - Estado da arte
  - Opiniões sobre o futuro da inteligência de máquina

Capítulo 2

- Caminhos para a superinteligência
  - Inteligência artificial
  - Emulação completa do cérebro
  - Cognição biológica
  - Interfaces cérebro-computador
  - Redes e organizações
  - Resumo

Capítulo 03

- Formas desuperinteligência
  - Superinteligência rápida

Superinteligência coletiva
Superinteligência de qualidade
Alcance direto e indireto
Vantagens da inteligência digital

Capítulo 04

A cinética de uma explosão de inteligência
Momento e velocidade da partida
Resistência
Caminhos de inteligência de não máquinas
Caminhos de emulação e IA
Poder de otimização e explosividade

Capítulo 05

Vantagem estratégica decisiva
O projeto líder terá uma vantagem estratégica decisiva?
Qual será o tamanho de um projeto bem-sucedido?
Monitoramento
Colaboração internacional
De uma vantagem estratégica decisiva para o singleton

Capítulo 06

Superpoderes cognitivos
Funcionalidades e superpoderes
Um cenário de tomada do poder por uma IA
Poder sobre a natureza e os agentes

Capítulo 07

A vontade superinteligente
A relação entre inteligência e motivação
Convergência instrumental
Autopreservação
Integridade do conteúdo do objetivo
Aprimoramento cognitivo
Perfeição tecnológica
Aquisição de recursos

Capítulo 08
O resultado mais provável será o nosso fim?
Catástrofe existencial como resultado esperado de uma explosão de inteligência?
A curva traiçoeira
Modos de falha malignos
Instanciação perversa
Profusão de infraestrutura
Crimes da mente

Capítulo 09
O problema do controle
Dois problemas de agência
Métodos de controle de capacidade
Métodos de confinamento
Métodos de incentivo
Inibição
Detonadores
Métodos de seleção de motivação
Especificação direta
Domesticidade

Normatividade indireta
Ampliação
Sinopse

Capítulo 10

Oráculos, gênios, soberanos e ferramentas
Oráculos
Gênios e soberanos
IAs ferramentas
Comparação

Capítulo 11

Cenários multipolares
Sobre cavalos e homens
Salários e desemprego
Capital e assistência social
O princípio malthusiano em uma perspectiva histórica
Crescimento populacional e investimento
A vida em uma economia algorítmica
Escravidão voluntária, morte casual
Trabalhar da forma mais eficiente possível seria divertido?
Inconscientes terceirizados?
Evolução não é necessariamente para cima
Formação de um singleton no pós-transição?
Uma segunda transição
Superorganismos e economias de escala
Unificação por tratado

Capítulo 12

Adquirindo valores

O problema da inserção de valores

Seleção evolutiva

Aprendizagem por reforço

Acréscimo associativo de valores

Estepe motivacional

Aprendizado de valores

Modulação de emulação

Projeto institucional

Sinopse

Capítulo 13

Elegendo os critérios de escolha

A necessidade da normatividade indireta

Vontade extrapolada coerente

Algumas explicações

Argumentos para a vec

Observações adicionais

Modelos de moralidade

Faça o que eu quero dizer

Lista de componentes

Conteúdo de objetivo

Teoria da decisão

Epistemologia

Ratificação

Chegando perto o suficiente

Capítulo 14

O panorama estratégico
Estratégia científica e tecnológica
Desenvolvimento tecnológico diferencial
Ordem preferível de chegada
Taxas de mudança e aperfeiçoamento cognitivo
Acoplamento de tecnologias
Antecipando reações
Caminhos e facilitadores
Os efeitos dos avanços em hardware
As pesquisas em emulação completa do cérebro deveriam ser promovidas?
A perspectiva do indivíduo-afetado favorece uma maior velocidade
Colaboração
A dinâmica competitiva e seus perigos
Sobre os benefícios da colaboração
Trabalhando em conjunto
Capítulo 15
A hora decisiva
Filosofia com prazo
O que deve ser feito?
Buscando a iluminação estratégica
Construindo uma boa capacidade
Medidas específicas
Que o melhor da natureza humana se coloque, por favor, de pé
Posfácio
Glossário parcial

Bibliografia

Agradecimentos

Sobre o autor

Créditos

SUPERINTELIGÊNCIA

FIGURAS, TABELAS E QUADROS

LISTA DE FIGURAS

01 Histórico do PIB mundial

02 Impacto geral de longo prazo da HLMI

03 Desempenho de supercomputadores

04 Reconstrução neuroanatômica em 3D a partir de imagens obtidas com um microscópio eletrônico

05 Roteiro para a emulação completa do cérebro

06 Composição de rostos como uma metáfora para a correção de genomas

07 A estrutura da partida

08 Uma escala menos antropomórfica?

09 Um modelo simples de uma explosão de inteligência

10 Fases em um cenário de tomada de poder por uma IA

11 Ilustração esquemática de trajetórias possíveis para um *singleton* prudente hipotético

12 Resultados da antropomorfização da motivação de um alienígena

13 Inteligência artificial ou emulação completa do cérebro primeiro?

14 Níveis de risco em corridas tecnológicas de IA

LISTA DE TABELAS

01 IA para jogos

02 Quando a inteligência de máquina de nível humano será alcançada?

03 Quanto tempo será necessário entre o nível humano e a superinteligência?

04 Capacidades necessárias para a emulação completa do cérebro

05 Ganhos máximos de QI a partir da seleção de embriões em conjuntos de diferentes tamanhos

06 Possíveis impactos da seleção genética em diferentes cenários

07 Algumas corridas tecnológicas estrategicamente significativas

08 Superpoderes: algumas tarefas estrategicamente relevantes e seus conjuntos de habilidades correspondentes

09 Diferentes tipos de detonadores

10 Métodos de controle

11 Características das diferentes castas de sistema

12 Sumário de técnicas de inserção de valores

13 Lista de componentes

LISTA DE QUADROS

01 Um agente bayesiano ótimo

02 O Flash Crash de 2010

03 O que seria necessário para replicar a evolução?

04 Sobre a cinética de uma explosão de inteligência

05 Corridas tecnológicas: alguns exemplos históricos

06 Cenário de encomenda de DNA

07 Qual é o tamanho do domínio cósmico?

08 Captura antrópica

09 Soluções estranhas em uma busca cega

10 Formalizando a aprendizagem de valores

11 Uma IA que deseja ser amigável

12 Duas ideias recentes (e pouco amadurecidas)

13 Um nivelamento por baixo arriscado

SUPER INTELI-GÊNCIA

PREFÁCIO

Dentro do seu crânio se encontra o responsável pela compreensão do mundo ao redor. O cérebro humano tem algumas capacidades inexistentes nos cérebros de outros animais. É a tais capacidades singulares que devemos nossa posição dominante no planeta. Outros animais possuem músculos mais fortes e garras mais afiadas, mas nós temos cérebros mais habilidosos. Nossa pequena vantagem em inteligência geral nos levou a desenvolver a linguagem, a tecnologia e uma organização social complexa. Essa vantagem tem aumentado ao longo do tempo, pois cada geração constrói em cima dos avanços conquistados por suas predecessoras.

Se algum dia construirmos cérebros artificiais capazes de superar o cérebro humano em inteligência geral, então essa nova superinteligência poderia se tornar muito poderosa. E, assim como o destino dos gorilas depende hoje mais dos humanos que dos próprios gorilas, também o destino de nossa espécie dependeria das ações da superinteligência de máquina.

Mas temos uma vantagem: nós é que construiríamos a máquina. Em princípio poderíamos construir um tipo de superinteligência que protegesse os valores humanos. Teríamos, certamente, fortes razões para isso. Na prática, o problema do controle — a questão de como controlar o que a superinteligência faria — parece bastante difícil. Tudo indica também que teríamos apenas uma única chance para resolvê-lo. Uma vez que uma superinteligência hostil passasse a existir, ela nos impediria de substituí-la ou de mudar suas preferências. Nosso destino estaria selado.

Neste livro, procuro entender o desafio que surge com a perspectiva da superinteligência e de que maneiras poderíamos melhor responder a ele. Esse é, muito provavelmente, o desafio mais importante e mais assustador que a humanidade já encarou. E, independentemente do nosso sucesso ou fracasso, promete ser o último desafio que encararemos.

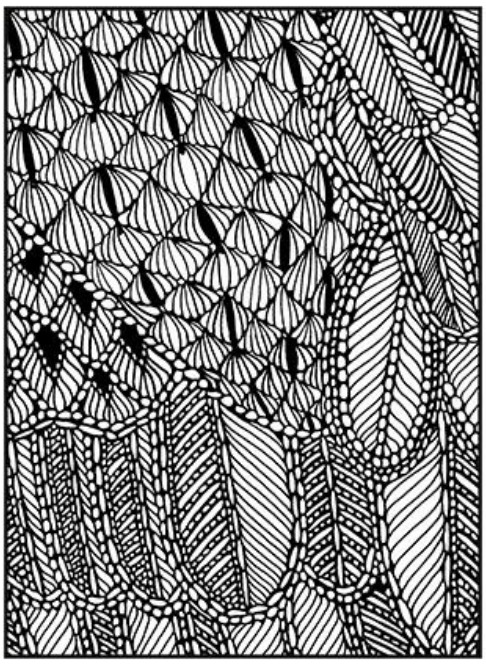
Este livro não argumenta que estejamos no limiar de uma grande descoberta em inteligência artificial, ou que possamos prever com alguma precisão quando um avanço desse tipo deve ocorrer. Mas é provável que isso aconteça ainda neste século, embora não tenhamos certeza disso. Os dois primeiros capítulos discutem possíveis caminhos e especulam sobre o momento mais provável para tal acontecimento. A maior parte do livro, no entanto, trata a respeito do que poderia acontecer depois disso. Estudamos a cinética de uma explosão da inteligência, as formas e poderes da superinteligência e as opções estratégicas disponíveis para um agente superinteligente que obtenha uma vantagem decisiva. Mais adiante, mudamos nosso foco para o problema do controle e questionamos o que poderia

ser feito para moldar as condições iniciais de modo a alcançar um resultado que garantisse nossa sobrevivência e nos trouxesse benefícios. Na parte final do livro, aumentamos nossa escala de observação e contemplamos o cenário maior que emerge a partir de nossas investigações. São fornecidas algumas sugestões sobre o que deve ser feito no presente para aumentar nossas chances de evitar uma catástrofe existencial no futuro.

Este não foi um livro fácil de escrever. Espero que o caminho aqui aberto permita que outros estudiosos alcancem essa nova fronteira da forma mais rápida e conveniente possível, a fim de que possam chegar lá menos desgastados e mais preparados para expandir ainda mais o alcance de nossa compreensão. (E se o processo é um pouco tortuoso, espero que os críticos, ao julgarem os resultados, não subestimem a hostilidade do terreno *ex ante*!)

Este não foi um livro fácil de escrever. Tentei escrever uma obra fácil de ler, mas não acredito que tenha sido bem-sucedido. Enquanto escrevia, imaginava um leitor que se parecesse comigo uns anos atrás e tentei produzir um livro cuja leitura me daria prazer. Seu público pode ser restrito. No entanto, acredito que o conteúdo deva ser acessível a muitas pessoas, se elas refletirem sobre ele e resistirem à tentação de interpretar erroneamente cada nova ideia, assimilando-a ao clichê mais parecido em sua bagagem cultural. Leitores sem familiaridade com os termos técnicos não devem se sentir desencorajados pela presença ocasional de fórmulas matemáticas ou pelo vocabulário especializado, pois é sempre possível entender a questão principal a partir das explicações apresentadas ao longo do texto. (E para os leitores que quiserem mais detalhes técnicos, há muitos deles nas notas de rodapé).[1]

Muitos dos argumentos apresentados neste livro estão provavelmente errados.[2] É possível também que haja considerações extremamente importantes que eu não tenha levado em conta, invalidando, dessa forma, partes ou mesmo o todo de minhas conclusões. Esforcei-me ao máximo para indicar as nuances e os níveis de incertezas do texto — poluindo-o com expressões tais como: “possivelmente”, “talvez”, “pode ser”, “me parece”, “deve ser”, “provavelmente”, “muito provavelmente”, “quase certamente”. Cada um desses termos foi posicionado de forma cuidadosa e deliberada. Ainda assim, essas aplicações pontuais de modéstia epistemológica não são suficientes, devendo ser aqui complementadas por uma confissão sistêmica de incerteza e falibilidade. Isto não é falsa modéstia: apesar de acreditar que meu livro possa estar completamente errado, apontando um caminho falso, penso que visões alternativas que vêm sendo apresentadas pela bibliografia especializada são ainda piores — incluindo aqui a visão padrão, ou “hipótese nula”, segundo a qual podemos por enquanto ignorar de forma segura e sensata a perspectiva da superinteligência.





01

CAPÍTULO 1

DESENVOLVIMENTOS ANTERIORES E CAPACIDADES ATUAIS

Começamos olhando para trás. A história, na sua escala mais ampla, parece apresentar uma sequência de distintos regimes de crescimento, cada um muito mais rápido que o seu predecessor. Esse padrão tem sido utilizado para sugerir que outro regime de crescimento (ainda mais rápido) pode ser possível. Entretanto, não colocamos muita ênfase nessa observação — este não é um livro que trata a respeito de "aceleração tecnológica", "crescimento exponencial" ou sobre as muitas noções algumas vezes reunidas sob a rubrica da "singularidade". Em seguida, repassamos a história da inteligência artificial. Mais adiante, examinamos as capacidades atuais desse campo de pesquisa. Por fim, analisamos algumas das pesquisas de opinião recentemente realizadas com especialistas e refletimos sobre nosso desconhecimento acerca da cronologia de futuros avanços.

Regimes de crescimento e a história

Há poucos milhões de anos, nossos ancestrais ainda estavam se balançando nos galhos das selvas africanas. Numa cronologia geológica ou mesmo evolucionária, o surgimento do *Homo sapiens* a partir do nosso último ancestral comum com os macacos aconteceu rapidamente. Desenvolvemos a postura ereta, os polegares opositores e — o mais importante — algumas mudanças relativamente pequenas no tamanho do nosso cérebro e na organização neurológica que resultaram num grande salto em nossa habilidade cognitiva. Como consequência, os humanos são capazes de pensar de forma abstrata, comunicar pensamentos complexos e acumular culturalmente informações através das gerações, de forma muito melhor do que qualquer outra espécie no planeta.

Essas capacidades levaram os humanos a desenvolver tecnologias de produção cada vez mais eficientes, tornando possível para nossos ancestrais a migração para áreas mais afastadas das florestas tropicais e das savanas. A densidade populacional e o tamanho total da população humana aumentaram, principalmente após a adoção da agricultura. Um número maior de pessoas levou a um número maior de ideias; uma maior densidade populacional permitiu que essas ideias se espalhassem mais rapidamente e que alguns indivíduos se dedicassem ao desenvolvimento de competências específicas. Esses desenvolvimentos aumentaram a *taxa de crescimento* da produtividade econômica e da capacidade tecnológica. Desenvolvimentos posteriores, relacionados à Revolução Industrial, provocaram um segundo salto de aumento na taxa de crescimento, comparável ao primeiro.

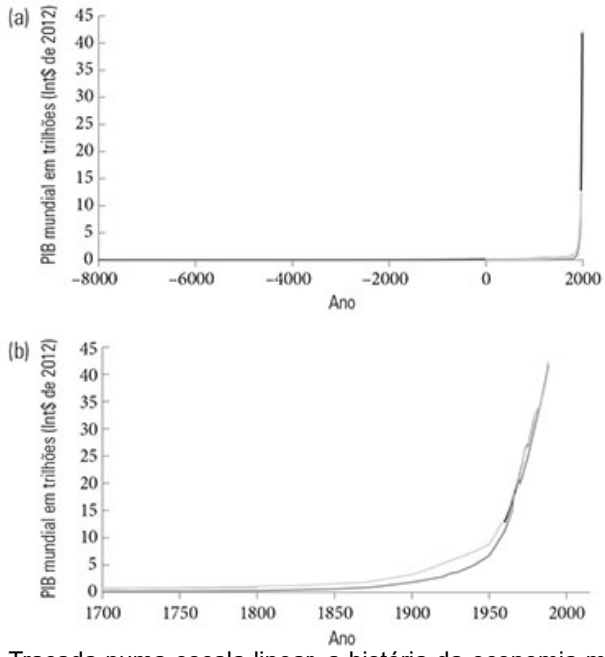
Tais mudanças na taxa de crescimento têm consequências importantes. Há poucas centenas de milhares de anos, no início da pré-história humana (ou hominídea), o crescimento era tão lento que foram necessários aproximadamente 1 milhão de anos até que a capacidade humana de produção aumentasse o suficiente para sustentar um adicional de 1 milhão de indivíduos vivendo no nível da subsistência. Em 5000 a.C., após a revolução agrícola, a taxa de crescimento havia aumentado a ponto de serem necessários apenas dois séculos para termos o mesmo volume de crescimento. Hoje, após a Revolução Industrial, a economia mundial cresce, em média, nesse mesmo volume a cada noventa minutos.[3]

Mesmo que a taxa de crescimento se mantenha nos níveis atuais, a longo prazo ela produzirá resultados impressionantes. Se a economia mundial continuar crescendo na mesma velocidade dos últimos cinquenta anos, o mundo será aproximadamente 4,8 vezes mais rico em 2050 e aproximadamente 34 vezes mais rico em 2100, em comparação ao que é hoje.[4]

Ainda assim, a perspectiva de continuar em um crescimento exponencial constante não se compara com o que poderia acontecer caso o mundo experimentasse novamente um salto na *taxa de crescimento*, comparável em magnitude àqueles experimentados com a Revolução Agrícola e a Revolução Industrial. O economista Robin Hanson estima, com base em dados histórico-econômicos e populacionais, que, para que a economia mundial dobrasse seu tamanho, tenham sido necessários 224 mil anos para as sociedades de caçadores-coletores do Plistoceno; 909 anos para as sociedades agrárias e 6,3 anos para a sociedade industrial.[5] (No modelo de Hanson, a época atual é uma mistura dos regimes de crescimento das sociedades agrária e industrial — a economia mundial como um todo ainda não está conseguindo dobrar sua taxa de crescimento a cada 6,3 anos.) Caso uma nova transição para um regime de crescimento diferente acontecesse, e esta fosse da mesma magnitude das duas anteriores, isso resultaria em um novo regime de crescimento no qual a economia mundial dobraria de tamanho a cada duas semanas, aproximadamente.

Uma taxa de crescimento dessa magnitude parece fantástica em comparação ao que conhecemos hoje. Observadores de épocas passadas talvez possam ter achado igualmente absurdo supor que a economia mundial duplicaria, no futuro, várias vezes ao longo da vida de uma pessoa. Porém, esse ritmo extraordinário é o que consideramos normal hoje em dia.

A ideia do surgimento de uma singularidade tecnológica tem sido largamente popularizada, começando com o artigo seminal de Vernor Vinge e seguindo com os escritos de Ray Kurzweil e outros.[6] O termo “singularidade”, no entanto, tem sido usado de forma confusa com sentidos distintos e vem acumulando uma aura profana (porém quase apocalíptica) com conotações tecnoutópicas.[7] Já que a maioria desses significados e conotações são irrelevantes para o nosso argumento, podemos ganhar clareza se dispensarmos o termo “singularidade”, optando por utilizar uma terminologia mais precisa.



HISTÓRIA DO PIB MUNDIAL. Traçada numa escala linear, a história da economia mundial aparece como uma linha plana que se mantém junto ao eixo × até que, de repente, apresenta um pico vertical. (a) Mesmo quando focamos nos últimos 10 mil anos, o padrão se mantém essencialmente como um único ângulo de 90o. (b) Apenas nos últimos cem anos, mais ou menos, a curva se eleva de forma perceptível acima do nível zero. (As diferentes linhas no gráfico correspondem a diferentes conjuntos de dados, que resultam em estimativas ligeiramente diferentes).*

**Van Zanden (2003); Maddison (1999, 2001); De Long (1998).*

A ideia relacionada à singularidade que nos interessa aqui é a possibilidade de uma *explosão de inteligência*, e particularmente a perspectiva de uma superinteligência de máquina. Há quem seja convencido pelos diagramas de crescimento, semelhantes ao da figura 1, de que uma nova mudança drástica do regime de crescimento comparável à Revolução Agrícola ou à Revolução Industrial esteja em vias de acontecer. Essas pessoas podem então achar difícil conceber um cenário no qual o tempo necessário para que a economia mundial dobre seu tamanho diminua para apenas algumas semanas sem que isso envolva a criação de mentes que sejam mais rápidas e eficientes do que as mentes biológicas que conhecemos. Entretanto, argumentos para levar a sério a perspectiva de uma revolução da inteligência artificial não precisam se basear em exercícios de ajuste de curva ou em extrapolações a partir do crescimento econômico experimentado no passado. Como veremos, há fortes razões para sermos cautelosos.

Grandes expectativas

Máquinas com inteligência geral comparável à dos humanos — ou seja, dotadas de bom senso e capacidade real de aprender, raciocinar e planejar a superação de desafios complexos de processamento de informação em uma vasta gama de domínios naturais e abstratos — têm sido esperadas desde a invenção dos computadores, na década de 1940. Naquele tempo, o advento de tais máquinas era frequentemente esperado para os vinte anos seguintes.[8] Desde então, a data

estimada para o seu surgimento tem recuado numa razão de um ano a cada ano, fazendo com que ainda hoje futuristas interessados na possibilidade de uma inteligência artificial geral acreditem que máquinas inteligentes surgirão dentro de duas décadas.[9]

Duas décadas é o período de tempo ideal para quem tenta prever uma mudança radical: próximo o suficiente para ser relevante e chamar atenção e ainda distante o suficiente para tornar possível imaginar que uma sequência de grandes avanços, atualmente apenas vagamente concebíveis, possa ocorrer. Contraponha isso a períodos de tempo menores: a maioria das tecnologias que nos próximos cinco ou dez anos terão um grande impacto no mundo já estão em uso de forma limitada, enquanto as tecnologias que remodelarão o mundo em menos de quinze anos provavelmente já existem em laboratórios na forma de protótipos. Vinte anos também pode ser um período próximo do tempo restante da carreira de quem trabalha prevendo tais avanços, limitando, assim, o risco que uma previsão ousada poderia causar à sua reputação.

O fato de que, no passado, alguns indivíduos fizeram previsões exageradas a respeito da inteligência artificial (IA) não implica, entretanto, que a IA seja impossível ou que nunca será desenvolvida.[10] A principal razão para que o progresso tenha sido mais lento do que o previsto se deve ao fato de que dificuldades técnicas para a construção de máquinas inteligentes se mostraram maiores do que os obstáculos previstos pelos pioneiros da IA. Mas isso não determina o tamanho dessas dificuldades e o quão longe estamos de superá-las. Muitas vezes, um problema que de início se mostra irremediavelmente complicado acaba tendo uma solução surpreendentemente simples (embora o oposto seja, provavelmente, mais comum).

No próximo capítulo, veremos os diferentes caminhos que podem nos levar à IA de nível humano. Mas notemos desde já que, independentemente de quantas paradas existam entre o presente e a IA de nível humano, ela não é o destino final. A próxima parada, a uma pequena distância adiante nos trilhos, é a IA sobre-humana. O trem pode não parar e nem mesmo desacelerar na Estação da Vila dos Humanos. É mais provável que ele passe direito.

O matemático I.J. Good, que tinha sido chefe de estatística da equipe de Alan Turing responsável pela quebra de códigos durante a Segunda Guerra Mundial, possivelmente foi o primeiro a enunciar os aspectos essenciais desse cenário. Em uma passagem datada de 1965, frequentemente citada, ele escreveu:

> Defina-se uma máquina ultrainteligente como uma máquina capaz de superar todas as atividades intelectuais de qualquer homem, independentemente de quão genial ele seja. Já que o projeto de máquinas é uma dessas atividades intelectuais, uma máquina ultrainteligente poderia projetar máquinas ainda melhores; haveria então certamente uma “explosão de inteligência”, e a inteligência humana se tornaria desnecessária. Desse modo, a primeira máquina ultrainteligente é a última invenção que o homem precisará fazer, contanto que a máquina seja dócil o suficiente para nos dizer como mantê-la sob controle.[11]

Pode parecer óbvio agora que grandes riscos existenciais estariam associados a essa explosão de inteligência, e que essa possibilidade deve ser examinada com a maior seriedade possível, mesmo se soubéssemos que a probabilidade de isso acontecer fosse pequena (o que não é o caso). Apesar da crença que os pioneiros da inteligência artificial tiveram na iminência de uma IA de nível humano, a maioria deles, no entanto, não concebeu a possibilidade de uma IA de nível superior ao humano. É como se toda a capacidade de especulação que eles possuíam tivesse se exaurido na concepção da possibilidade radical de máquinas alcançando uma inteligência humana e, portanto, não foram capazes de compreender o corolário — que as máquinas se tornariam, posteriormente, superinteligentes.

A maioria dos pioneiros de IA não consideraram a possibilidade de que sua empreitada poderia envolver alguns riscos.[12] Eles não fizeram declarações retóricas — muito menos considerações sérias — sobre qualquer preocupação referente à segurança ou questões éticas relacionadas à criação de mentes artificiais e de potenciais computadores déspotas: uma lacuna que surpreende mesmo diante do padrão não-tão-impressionante das avaliações críticas de tecnologia da época.[13] Devemos ter a esperança de que, quando a empreitada finalmente se tornar viável, nós tenhamos alcançado não apenas a proficiência tecnológica para desencadear uma explosão de inteligência, mas também o nível maior de maestria necessário para que sejamos capazes de sobreviver à detonação dessa explosão.

Mas antes que nos voltemos para o que vem adiante, será útil darmos uma olhada na história da inteligência artificial até o presente momento.

Períodos de esperança e aflição

No verão de 1956, na Dartmouth College, dez cientistas que compartilhavam o interesse em redes neurais, na teoria dos autômatos e no estudo da inteligência se reuniram para um workshop de seis semanas. O Dartmouth Summer Project é considerado, frequentemente, o marco que assinalou a inteligência artificial como um campo de pesquisa. Muitos dos participantes seriam mais tarde reconhecidos como fundadores desse campo. A perspectiva otimista dos participantes está refletida na proposta apresentada à Rockefeller Foundation, que patrocinou o evento:

> Nós propomos que um estudo sobre inteligência artificial seja realizado por dez homens durante duas semanas [...]. O estudo será realizado com base na hipótese de que todos os aspectos da aprendizagem ou de qualquer outra característica da inteligência podem, em princípio, ser descritos de forma tão precisa que uma máquina poderia simulá-los. Será feita uma tentativa de descobrir como produzir máquinas que usem linguagem, formem conceitos e abstrações, resolvam determinados tipos de problemas até agora reservados aos humanos e que sejam capazes de se aperfeiçoar. Acreditamos que um avanço significativo possa ser feito em um ou mais desses problemas se um

> grupo de cientistas cuidadosamente selecionados trabalhar em conjunto no projeto durante um verão.

Nas seis décadas seguintes a esse início audacioso, o campo da inteligência artificial passou por períodos de expectativas exageradas alternados com períodos de recuos e decepções.

O primeiro período de entusiasmo, que iniciou com o encontro em Dartmouth, foi mais tarde descrito por John McCarthy (o principal organizador do evento) como a era do "Veja, mamãe, sem as mãos!". Nessa época, pesquisadores construíram sistemas projetados para contestar afirmações como: "Nenhuma máquina jamais seria capaz de fazer X!". Tais afirmações céticas eram comuns no período. Para contrapô-las, os pesquisadores da IA criavam pequenos sistemas que faziam X em um "micromundo" (um domínio bem definido e limitado que tornava possível uma versão reduzida da performance a ser demonstrada), fornecendo, dessa forma, uma comprovação do conceito e mostrando que X poderia, em princípio, ser feito por uma máquina. Um desses sistemas iniciais, o Logic Theorist, foi capaz de provar a maioria dos teoremas que constam no segundo capítulo do *Principia Mathematica*, de Whitehead e Russel, e produziu até mesmo uma prova que era muito mais elegante que a original, desmascarando, dessa forma, a noção de que as máquinas poderiam "pensar apenas numericamente" e mostrando que as máquinas eram também capazes de fazer deduções e inventar provas lógicas.[14] Um sucessor desse sistema, o General Problem Solver, poderia, em princípio, resolver uma vasta gama de problemas definidos formalmente.[15] Programas capazes de resolver problemas de cálculo típicos de primeiro ano de cursos de graduação, problemas de analogia visual que aparecem em alguns testes de QI e problemas simples de álgebra em linguagem verbal também foram escritos.[16] O robô Shakey (assim chamado em virtude da sua tendência de tremer durante o seu funcionamento) demonstrou como o raciocínio lógico poderia ser integrado à percepção e usado para planejar e controlar atividades físicas.[17] O programa eliza mostrou como um computador poderia imitar um psicoterapeuta rogeriano.[18] Em meados da década de 1970, o programa shrdlu demonstrou como a simulação de um braço robótico num mundo simulado de blocos geométricos poderia seguir instruções e responder questões em inglês que eram digitadas por um usuário.[19] Nas décadas seguintes, outros sistemas foram criados e comprovaram que as máquinas eram capazes de compor músicas no estilo de vários compositores clássicos, superar médicos recém-formados em certas tarefas relacionadas a diagnósticos clínicos, dirigir carros de forma autônoma e criar invenções patenteáveis.[20] Há ainda uma IA capaz de contar piadas originais.[21] (Não que o nível do humor seja elevado, mas crianças relataram achar seus trocadilhos divertidos.)

Os métodos que produziram sucessos nos primeiros sistemas de demonstração geralmente apresentaram dificuldades de extensão para uma variedade mais ampla de problemas ou para versões mais complexas do problema. Uma razão para isso é a "explosão combinatorial" de possibilidades que devem ser exploradas por métodos que se baseiam em algo como a busca exaustiva. Tais métodos funcionam

bem para instâncias simples de um problema, mas falham quando as coisas ficam um pouco mais complicadas. Por exemplo, para provar um teorema que tenha uma prova de cinco linhas num sistema de dedução com uma regra de inferência e cinco axiomas, alguém poderia simplesmente enumerar as 3.125 combinações possíveis e conferir cada uma delas para checar sua correspondência com a conclusão pretendida. A busca exaustiva também funcionaria para provas de seis ou sete linhas. Mas quando as tarefas ficam mais difíceis, o método de busca exaustiva logo se torna inviável. Provar um teorema com uma prova de cinquenta linhas não exige dez vezes o tempo necessário para uma prova de cinco linhas: na verdade, o uso da busca exaustiva requer a varredura de $5^{50} \approx 8{,}9 \times 10^{34}$ sequências possíveis — o que é computacionalmente inviável mesmo com os supercomputadores mais rápidos.

Para superar a explosão combinatorial, são necessários algoritmos que explorem a estrutura no domínio alvo e tirem vantagem de conhecimento prévio usando busca heurística, planejamento e representações abstratas flexíveis — capacidades que ainda estavam pouco desenvolvidas nos sistemas iniciais de IA. O fraco desempenho desses primeiros sistemas também se devia a diversos outros fatores, tais como: métodos inadequados para lidar com incertezas, dependência de representações simbólicas frágeis e infundadas, escassez de dados e severas limitações de hardware tanto em capacidade de memória quanto em velocidade de processamento. Em meados da década de 1970, já havia uma conscientização crescente acerca desses problemas. A constatação de que muitos dos projetos de IA jamais poderiam realizar suas promessas iniciais levou ao surgimento do primeiro "inverno da IA": um período de retração durante o qual os financiamentos diminuíram, o ceticismo aumentou, e a IA, então, saiu de moda.

Uma nova primavera chegou no início dos anos 1980, quando o Japão lançou o seu projeto de sistemas de computador de quinta geração, uma parceria público-privada com ótimo financiamento cujo objetivo era ultrapassar o estado da arte desenvolvendo uma arquitetura computacional maciçamente paralela que serviria como uma plataforma para a inteligência artificial. Isso ocorreu no auge do fascínio com o "milagre econômico japonês do pós-guerra", um período em que governos e grandes empresários ocidentais procuravam ansiosamente adivinhar a fórmula por trás do sucesso econômico japonês na esperança de replicar a mágica em casa. Quando o Japão decidiu investir pesado em IA, vários outros países resolveram segui-lo.

Nos anos seguintes aconteceu uma grande proliferação de *sistemas especialistas*. Projetados como ferramentas de suporte para tomadas de decisão, os sistemas especialistas eram programas baseados em regras que faziam inferências simples a partir de uma base de conhecimento composta de fatos extraídos de especialistas na área e programados manualmente em linguagem formal, e de forma trabalhosa. Centenas desses sistemas especialistas foram construídos. No entanto, os sistemas menores forneciam poucos benefícios, e os maiores se mostraram caros demais para serem desenvolvidos, validados e atualizados, além de serem difíceis de usar. Era impraticável adquirir um

computador apenas para rodar um único programa. No fim da década de 1980, essa temporada de crescimento também encerrou seu ciclo.

O projeto da .quinta geração fracassou em alcançar seus objetivos, assim como fracassaram seus similares nos Estados Unidos e na Europa. Surgiu, então, um segundo inverno na IA. Naquele momento, um crítico poderia justificadamente lamentar que "a história da pesquisa em inteligência artificial até hoje sempre consistiu de sucessos muito limitados em áreas específicas, seguidos imediatamente pelo fracasso em alcançar os objetivos maiores sugeridos, à primeira vista, por esses sucessos iniciais".[22] Investidores do setor privado começaram a evitar qualquer empreendimento que carregasse a marca da "inteligência artificial". Mesmo entre os acadêmicos e seus financiadores, a "IA" se tornou um epíteto indesejável.[23]

No entanto, o trabalho técnico se manteve num ritmo acelerado e, nos anos 1990, o segundo inverno da IA teve seu degelo. O otimismo renasceu a partir da introdução de novas técnicas que pareciam oferecer alternativas ao paradigma lógico tradicional (geralmente chamado de *Good Old-Fashioned Artificial Intelligence* [boa e velha inteligência artificial], ou usando a sigla GOFAI), o qual havia se concentrado na manipulação de símbolos de alto nível e que atingiu seu apogeu com os sistemas especialistas da década de 1980. As técnicas que ficaram mais conhecidas nessa época, que incluíam redes neurais e algoritmos genéticos, prometeram superar algumas limitações da abordagem GOFAI, em particular a "fragilidade" que caracterizava os programas clássicos de IA (que normalmente geravam resultados sem nenhum sentido caso fossem programados com uma única hipótese ligeiramente errada). As novas técnicas exibiam um desempenho mais orgânico. Por exemplo, redes neurais exibiam a propriedade da "degradação graciosa": um pequeno dano a uma rede neural normalmente resultava numa pequena degradação de sua performance, em vez de um colapso total. Ainda mais importante: redes neurais poderiam aprender a partir de experiências, encontrando caminhos naturais de generalização por meio de exemplos e padrões estatísticos ocultos nos dados de entrada.[24] Isso fez com que as redes tivessem bom desempenho em reconhecimento de padrões e problemas de classificação. Por exemplo, ao treinar uma rede neural num conjunto de dados de sinais de sonar, ela poderia ser ensinada a diferenciar os perfis acústicos de submarinos, minas e vida marinha com mais exatidão que os especialistas humanos — e isso poderia ser feito sem que alguém tivesse que descobrir com antecedência como as categorias seriam definidas ou como as diferentes características teriam que ser ponderadas exatamente.

Embora modelos simples de rede neural sejam conhecidos desde o fim da década de 1950, o campo de pesquisa experimentou um renascimento após a introdução do algoritmo de retropropagação (*backpropagation*), que permitiu treinar redes neurais multicamadas.[25] Tais redes multicamadas, que possuem uma ou mais camadas intermediárias ("ocultas") de neurônios entre as camadas de entrada e saída de dados, podem aprender uma gama muito mais variada de funções do que seus predecessores mais simples.[26] Combinados com os computadores cada vez mais poderosos que estavam se tornando disponíveis, esses aperfeiçoamentos

algorítmicos capacitaram engenheiros na construção de redes neurais boas o suficiente para serem úteis, na prática, em várias aplicações.

As qualidades semelhantes às do cérebro que as redes neurais exibiam contrastavam favoravelmente com o detalhismo lógico rígido, mas frágil, dos sistemas GOFAI baseados em regras. Isso foi suficiente para inspirar um novo "ismo", o "conexionismo", que enfatizava a importância do processamento subsimbólico maciçamente paralelo. Mais de 150 mil artigos acadêmicos sobre redes neurais artificiais foram publicados desde então, e elas continuam a ser uma abordagem importante do aprendizado de máquina.

Métodos com base no evolucionismo, tais como os algoritmos genéticos e a programação genética, constituem outra abordagem cujo surgimento ajudou a acabar com o segundo inverno da IA. Talvez essa abordagem tenha causado um impacto acadêmico menor do que o das redes neurais, mas foi largamente popularizada. Em modelos evolucionários, uma população de soluções candidatas (que podem ser programas ou estruturas de dados) é mantida e novas soluções candidatas são geradas aleatoriamente através de mutação ou recombinação de variáveis da população existente. Periodicamente a população é podada por meio da aplicação de um critério de seleção (função de aptidão), permitindo que apenas os melhores candidatos sobrevivam na geração seguinte. Quando esse processo é repetido por milhares de gerações, a média da qualidade das soluções candidatas aumenta gradualmente. Quando funciona, esse tipo de algoritmo consegue produzir soluções eficientes para uma gama altamente variada de problemas — soluções que podem ser surpreendentemente inovadoras e não intuitivas, geralmente se parecendo mais com estruturas naturais do que qualquer coisa que um engenheiro humano projetaria. E, em princípio, isso pode acontecer sem muita necessidade de contribuição humana além da especificação inicial da função de aptidão, o que geralmente é muito simples. Na prática, entretanto, conseguir que métodos evolucionários funcionem bem requer habilidade e criatividade, particularmente para elaborar boas formas de representação. Sem uma maneira eficiente de codificar soluções candidatas (uma linguagem genética que esteja de acordo com a estrutura latente no domínio em questão), a busca evolucionária tende a vagar indefinidamente em um vasto espaço de busca ou a ficar presa em um ótimo local. Mesmo que uma boa forma de representação seja encontrada, a evolução exige muitos recursos computacionais e é frequentemente derrotada pela explosão combinatorial.

Redes neurais e algoritmos genéticos são exemplos de métodos que causaram empolgação na década de 1990 ao sugerir que ofereciam alternativas para a estagnação do paradigma GOFAI. Mas a intenção aqui não é a de defender esses dois métodos ou de elevá-los acima de outras técnicas de aprendizado de máquina. Na verdade, um dos maiores desenvolvimentos teóricos dos últimos vinte anos foi a constatação clara de que técnicas superficialmente distintas podem ser entendidas como casos especiais dentro de uma estrutura matemática comum. Por exemplo, muitos tipos de redes neurais artificiais podem ser vistas como classificadores que realizam um tipo específico de cálculo estatístico (estimativa de máxima verossimilhança).[27] Essa perspectiva permite que as redes neurais sejam

comparadas com uma classe maior de algoritmos para aprendizado de classificadores a partir de exemplos — “árvores de decisão”, “modelos de regressão logística”, “máquinas de vetores de suporte”, “classificador naive Bayes”, “regressão de K-vizinhos mais próximos”, entre outros.[28] De modo semelhante, algoritmos genéticos podem ser vistos como algoritmos de busca que executam “subida de encosta estocástica” e que são, por sua vez, parte de uma categoria mais ampla de algoritmos de otimização. Cada um desses algoritmos utilizados na construção de classificadores ou para fazer busca em um espaço de soluções tem seu próprio perfil com vantagens e desvantagens que podem ser estudadas matematicamente. Os algoritmos se distinguem pelos requisitos de tempo de processamento e espaço de memória, pelos vieses indutivos que eles pressupõem, pela facilidade com a qual conseguem incorporar conteúdo produzido externamente e por quão transparente seu funcionamento interno se mostra para um analista humano.

Logo, por trás do alvoroço que envolve as áreas de aprendizado de máquina e resolução criativa de problemas, encontra-se uma série de escolhas com perdas e ganhos matematicamente bem definidos. O ideal em questão é o do agente bayesiano perfeito, que faz uso probabilisticamente ótimo das informações disponíveis. Esse ideal é inalcançável, pois tem uma demanda computacional muito grande para ser implementado em qualquer computador físico (ver quadro 1). Assim, é possível ver a inteligência artificial como uma procura por atalhos: meios de se aproximar do ideal bayesiano sacrificando alguma otimização ou generalização, enquanto se preserva o suficiente para atingir uma alta performance nos reais domínios de interesse.

Um reflexo desse cenário pode ser visto no trabalho realizado nas últimas duas décadas em modelos gráficos probabilísticos, tais como as redes bayesianas. Redes bayesianas oferecem uma forma concisa de se representar relações de independência probabilística e condicional válidas para alguns domínios específicos. (Explorar tais relações de independência é essencial para superar a explosão combinatorial, que é um grande problema, tanto para a inferência probabilística quanto para a dedução lógica.) Eles também ajudam na compreensão do conceito de causalidade.[29]

Uma vantagem de se relacionar problemas de aprendizagem oriundos de domínios específicos ao problema geral de inferência bayesiana é que novos algoritmos que tornem a inferência bayesiana mais eficiente permitirão, dessa forma, melhorias imediatas em várias áreas diferentes. Avanços em técnicas de aproximação de Monte Carlo, por exemplo, são diretamente aplicáveis em visão computacional, robótica e genética computacional.

QUADRO 1

Um agente bayesiano ótimo

Um agente bayesiano ideal começa com uma “distribuição de probabilidade *a priori*”, uma função que atribui probabilidades para cada “mundo possível” (ou seja, para cada descrição inteiramente especificada de como o mundo poderia ser na realidade).[30] Esta função incorpora um viés indutivo que faz com que probabilidades maiores sejam designadas a mundos possíveis mais simples. (Uma maneira de definir formalmente a simplicidade de um mundo possível é em termos de sua “complexidade de Kolmogorov”, uma medida baseada no tamanho do menor programa de computador que gera uma descrição completa do mundo).[31] A

distribuição de probabilidade *a priori* também incorpora qualquer conhecimento prévio que os programadores desejarem dar ao agente.

Assim que o agente recebe uma nova informação dos seus sensores, ele atualiza sua distribuição de probabilidade condicionando a distribuição à nova informação de acordo com o teorema de Bayes.[32] Condicionalização é a operação matemática que modifica para zero a nova probabilidade dos mundos que são inconsistentes com a informação recebida e renormaliza a distribuição de probabilidade considerando apenas os mundos possíveis restantes. O resultado é uma "distribuição de probabilidade a posteriori" (que o agente pode utilizar como sua distribuição de probabilidade *a priori* na próxima etapa). Enquanto o agente faz suas observações, sua massa de probabilidade se concentra numa série decrescente de mundos possíveis que permanecem consistentes com a evidência, sendo que, entre esses mundos possíveis, os mais simples sempre têm maior probabilidade.

Metaforicamente, podemos pensar em probabilidade como uma quantidade de areia sobre uma grande folha de papel. O papel é dividido em áreas de vários tamanhos, cada área correspondendo a um mundo possível, e áreas maiores representando mundos possíveis mais simples. Imagine também uma camada de areia com espessura uniforme espalhada por toda a folha: essa é nossa distribuição de probabilidade *a priori*. Toda vez que uma observação feita descarta alguns mundos possíveis, nós removemos a areia correspondente a essas áreas do papel e a redistribuímos igualmente sobre as áreas que permanecem em jogo. Dessa forma, a quantidade total de areia sobre a folha nunca muda, mas apenas se concentra em menos áreas à medida que a evidência observacional se acumula. Esse é um retrato do aprendizado na sua forma mais pura. (Para calcular a probabilidade de uma hipótese, simplesmente medimos a quantidade de areia em todas as áreas que correspondem aos mundos possíveis nos quais a hipótese é verdadeira.)

Até o momento, definimos uma regra de aprendizado. Para obter um agente, precisamos também de uma regra de decisão. Para tal fim, dotamos o agente com uma "função de utilidade" que atribui um número a cada mundo possível. O número representa a conveniência daquele mundo de acordo com as preferências básicas do agente. A partir de então, a cada passo, o agente seleciona a ação com a maior utilidade esperada.[33] (Para encontrar a ação com a maior utilidade esperada, o agente poderia listar todas as possíveis ações. Ele poderia então calcular a distribuição de probabilidade condicional dada a ação — a distribuição de probabilidade que resultaria do condicionamento de sua distribuição de probabilidade atual à observação de que a ação acabou de ser tomada. Por fim, ele calcularia o valor esperado da ação como a soma do valor de cada mundo possível multiplicado pela probabilidade condicional daquele mundo dada a ação.)[34]

A regra de aprendizado e a regra de decisão, juntas, definem uma "noção de otimalidade" para um agente. (Essencialmente, a mesma noção de otimalidade tem sido amplamente utilizada em inteligência artificial, epistemologia, filosofia da ciência, economia e estatística).[35] Na realidade, é impossível construir tal agente por ser computacionalmente intratável realizar os cálculos necessários. Qualquer tentativa de fazê-lo sucumbe a uma explosão combinatória exatamente como a que descrevemos em nossa discussão a respeito da GOFAI. Para entender porque isso acontece, considere um pequeno subconjunto de todos os mundos possíveis: aqueles que consistem de um único monitor de computador flutuando num vácuo infinito. O monitor tem 1.000 × 1.000 pixels, dos quais cada qual está permanentemente ligado ou desligado. Mesmo esse subgrupo de mundos possíveis é imensamente grande: as $2^{(1.000 \times 1.000)}$ possibilidades de estado do monitor superam todas as computações esperadas para acontecer no universo observável. Dessa forma, não poderíamos nem ao menos enumerar todos os mundos possíveis nesse pequeno subconjunto de todos os mundos possíveis, quem dirá, então, executar cálculos mais elaborados em cada um deles, individualmente.

Noções de otimalidade podem ser de interesse teórico mesmo que sejam fisicamente inconcebíveis. Elas nos fornecem um padrão com o qual podemos julgar aproximações heurísticas, e em alguns casos é possível raciocinar a respeito do que um agente ótimo faria em algum caso especial. Encontraremos algumas outras noções de otimalidade para agentes artificiais no capítulo 12.

TABELA 1

IA para jogos

Damas	Super-humana	O programa de damas de Arthur Samuel, originalmente escrito em 1952 e mais tarde aprimorado (a versão de 1955 incorporou o aprendizado de máquina), tornou-se o primeiro programa a aprender a jogar melhor um jogo do que o seu criador.[36] Em 1994, o programa CHINOOK derrotou o campeão humano da época, marcando a primeira vez que um programa ganhou um campeonato mundial oficial num jogo de

		habilidade. Em 2002, Jonathan Schaeffer e sua equipe "solucionaram" o jogo de damas, ou seja, produziram um programa que sempre faz a melhor jogada possível (combinando busca alfa-beta com um banco de dados de 39 trilhões de posições finais de jogo). Um jogo perfeito de ambos os lados leva ao empate.[37]
Gamão	Super-humana	1979: o programa de gamão BKG, de Hans Berliner, derrotou o campeão mundial — o primeiro programa de computador a derrotar (em uma partida amistosa) um campeão mundial em qualquer jogo —, apesar de Berliner mais tarde ter atribuído a vitória à sorte com os dados.[38] 1992: O programa de gamão TD-Gammon, de Gerry Tesauro, alcançou uma habilidade de nível de competição, usando aprendizado de diferença temporal (uma forma de aprendizagem por reforço) e jogos repetidos contra ele mesmo para se aperfeiçoar.[39] Nos anos seguintes, outros programas de gamão superaram, em muito, os melhores jogadores humanos.[40]
Traveller TCS	Super-humana em colaboração com humanos[6]	Em 1981 e 1982[41], o programa Eurisko, de Douglas Lenat, vence o campeonato dos Estados Unidos de Traveller TCS (um jogo de guerra naval futurista) induzindo mudanças de regras para bloquear suas estratégias não ortodoxas.[42] O programa Eurisko tinha uma heurística para projetar suas esquadras e também uma heurística para modificar sua heurística.
Othello	Super-humano	1997: o programa Logistello ganhou todas as partidas numa série de seis jogos contra o campeão mundial Takeshi Murakami.[43]
Xadrez	Super-humana	1997: Deep Blue vence o campeão mundial de xadrez, Garry Kasparov. Kasparov diz ter visto lampejos de inteligência e criatividade verdadeiras em algumas das jogadas do computador.[44] Desde então, as máquinas de xadrez continuaram se aperfeiçoando.[45]
Palavras Cruzadas	Nível de especialista	1999: o Proverb, um programa de solução de palavras cruzadas, supera um solucionador humano de palavras cruzadas de nível médio.[46] 2012: o programa Dr. Fill, criado por Matt Ginsberg, fica no melhor quartil entre os competidores humanos no American Crossword Puzzle Tournament. (A performance do Dr. Fill foi irregular. Ele completou perfeitamente o quebra-cabeça considerado como o mais difícil pelos humanos, mas teve dificuldade de completar alguns quebra-cabeças fora do padrão, com palavras soletradas de trás para a frente ou na diagonal.)[47]
Scrabble	Super-humano	Desde 2002, softwares superam os melhores jogadores humanos.[48]
Bridge	Igual ao melhor humano	Em 2005, o software para jogar brigde alcança a paridade com os melhores jogadores humanos de bridge.[49]
Jeopardy!	Super-humana	2010: O *Watson* da IBM derrota os dois maiores campeões mundiais de *Jeopardy!*, Ken Jennings e Brad Rutter.[50] *Jeopardy!* é um jogo de perguntas sobre história, literatura, esportes, geografia, cultura popular, ciência e outros tópicos, transmitido na televisão nos Estados Unidos. As questões são apresentadas em forma de dicas e geralmente envolvem trocadilhos.

Pôquer	Variada	Computadores jogadores de pôquer continuam levemente abaixo dos melhores humanos na modalidade *Texas hold'em* com mesa de apostas cheia, mas atuam em nível super-humano em algumas variantes do jogo.[51]
FreeCell	Super-humana	Heurísticas evoluídas com o uso de algoritmos genéticos produziram um solucionador para o jogo FreeCell (o qual, em sua forma generalizada, é NP-completo) que é capaz de vencer jogadores humanos de alto nível.[52]
Go	Ótimo nível amador	Em 2012, a série Zen de programas jogadores de Go alcançou a categoria de sexto *dan* em jogos rápidos (nível de um excelente jogador amador), usando busca Monte Carlo em árvores de hipóteses e técnicas de aprendizado de máquina.[53] Programas jogadores de Go têm sido aperfeiçoados numa taxa de um *dan* por ano nos últimos anos. Se essa média de aperfeiçoamento continuar, eles poderão vencer o campeão mundial humano em cerca de uma década.

Outra vantagem é permitir que pesquisadores de diferentes disciplinas agreguem mais facilmente suas descobertas. Modelos gráficos e estatística bayesiana têm se tornado um foco comum de pesquisa em muitos campos, incluindo aprendizado de máquina, física estatística, bioinformática, otimização combinatória e teoria da comunicação.[54] Uma quantidade considerável do progresso recente em aprendizado de máquina é produto da incorporação de resultados formais obtidos originalmente em outros campos acadêmicos. (Aplicações de aprendizado de máquina também se beneficiaram enormemente de computadores mais rápidos e de uma maior disponibilidade de grandes conjuntos de dados.)

Estado da arte

A inteligência artificial já supera a inteligência humana em vários domínios. A tabela 1 examina o estágio atual da inteligência artificial para jogos, mostrando que IAs atualmente superam campeões humanos em uma gama variada de jogos.[55]

Essas conquistas podem não nos impressionar hoje. Mas isso é porque adaptamos nosso critério do que é impressionante aos avanços já alcançados. A maestria no xadrez já foi considerada o epítome do intelecto humano. Na opinião de vários especialistas no fim dos anos 1950: "Se alguém pudesse elaborar uma máquina capaz de jogar xadrez, essa pessoa teria, supostamente, penetrado no âmago da empreitada intelectual humana".[56] Mas, atualmente, isso não é mais verdade. Há quem concorde com John McCarthy, que lamentou: "Quando algo funciona, ninguém mais chama isso de IA".[57]

Há um sentido importante, no entanto, no qual as IAs jogadoras de xadrez se mostraram um triunfo menor do que muitos imaginavam. Anteriormente se supunha, não totalmente sem razão, que para um computador jogar xadrez no nível de um grande mestre, teria que ser dotado com um alto grau de inteligência *geral*.[58] Imaginava-se, por exemplo, que jogar xadrez requeria capacidade de aprender conceitos abstratos, pensar de forma inteligente sobre estratégia, compor planos flexíveis, empregar uma ampla gama de deduções lógicas complexas e talvez até

mesmo modelar o pensamento do seu oponente. Mas isso não é verdade. Confirmou-se que é possível construir um mecanismo de xadrez perfeitamente bom em torno de um algoritmo de propósito específico.[59] Quando esse algoritmo foi implementado nos processadores rápidos que se tornaram disponíveis no fim do século xx, resultou num excelente jogo. Mas uma IA construída dessa forma é limitada. Ela não faz nada além de jogar xadrez.[60]

Em outros domínios de aplicação, soluções têm se mostrado *mais* complicadas do que se esperava inicialmente e progridem muito mais lentamente. O cientista da computação Donald Knuth observou que "a IA tem tido sucesso até agora em fazer essencialmente tudo o que requer 'pensar', mas tem falhado em fazer a maior parte daquilo que as pessoas e os animais fazem 'sem pensar' — isso, de alguma forma, é muito mais difícil!".[61] Analisar cenas visuais, reconhecer objetos ou controlar o comportamento de um robô enquanto ele interage com um ambiente natural se provaram problemas desafiadores. Entretanto, um progresso considerável foi e continua a ser alcançado, auxiliado pelo aperfeiçoamento constante dos hardwares de computadores.

Senso comum e compreensão de linguagem natural também têm se mostrado difíceis. Considera-se com frequência que atingir um desempenho de nível plenamente humano nessas tarefas é um problema "IA-completo", significando que a dificuldade de solucionar esses problemas se equivale à dificuldade de se construir máquinas com inteligência geral de nível humano.[62] Em outras palavras, se alguém *tivesse* sucesso na criação de uma IA que entendesse a linguagem natural tanto quanto um humano adulto, esse alguém muito provavelmente já teria logrado sucesso na criação de uma IA capaz de fazer tudo mais que a inteligência humana pode fazer, ou estaria muito próximo disso.[63]

A proficiência no jogo de xadrez se mostrou alcançável por meio de um algoritmo surpreendentemente simples. É tentador especular que outras capacidades — tais como a habilidade geral de raciocínio, ou alguma habilidade específica de programação — podem igualmente ser alcançadas através de algum algoritmo surpreendentemente simples. O fato de que o melhor desempenho é obtido uma vez por meio de um mecanismo complicado não significa que um mecanismo simples não poderia fazer a tarefa tão bem quanto ou ainda melhor. Pode ser que simplesmente não se tenha ainda encontrado a alternativa mais simples. O sistema ptolemaico (com a Terra ao centro, orbitada pelo Sol, a Lua, os planetas e as estrelas) representou o estado da arte em astronomia por mais de mil anos, e sua precisão preditiva foi aprimorada ao longo dos séculos através de modificações que tornavam o modelo cada vez mais complexo: adicionando epiciclos sobre epiciclos aos movimentos celestiais postulados. Então, todo o sistema foi derrubado pela teoria heliocêntrica de Copérnico, que era mais simples e — embora apenas após elaborações posteriores de Kepler — fazia previsões mais precisas.[64]

Métodos de inteligência artificial são agora utilizados em mais áreas do que faria sentido revisar aqui, mas mencionar uma amostra delas nos dará uma ideia da abrangência de suas aplicações. Além das IAs para jogos listadas na tabela 1, há aparelhos auditivos com algoritmos que filtram ruído ambiente; traçadores de rota que mostram mapas e oferecem conselhos de navegação aos motoristas; sistemas

de recomendação que sugerem livros e álbuns de música baseados nas compras prévias e avaliações do usuário; sistemas de apoio a decisões médicas que ajudam os profissionais a diagnosticarem câncer de mama, recomendam planos de tratamento e auxiliam na interpretação de eletrocardiogramas. Há robôs de estimação e robôs de limpeza, robôs cortadores de grama, robôs para resgates, robôs cirurgiões e mais de 1 milhão de robôs na indústria.[65] A população mundial de robôs excede o número de 10 milhões.[66]

Programas de reconhecimento de voz modernos, baseados em técnicas estatísticas como modelos ocultos de Markov, têm se tornado precisos o suficiente para o uso prático (algumas partes deste livro foram redigidas com a ajuda de um programa do tipo). Assistentes digitais pessoais, como a Siri, da Apple, reagem a comandos falados e podem responder a perguntas simples e executar comandos. Programas de reconhecimento óptico de caracteres de textos escritos à mão ou à máquina são utilizados rotineiramente em aplicativos de triagem de correspondência e de digitalização de documentos antigos.[67]

Traduções feitas por máquinas continuam imperfeitas, mas boas o suficiente em muitas aplicações. Os primeiros sistemas utilizaram a abordagem GOFAI de gramáticas manualmente programadas que foram desenvolvidas do zero por linguistas especializados para cada idioma específico. Sistemas mais novos utilizam técnicas estatísticas de aprendizado de máquina que automaticamente constroem modelos estatísticos a partir de padrões de uso observados. A máquina infere os parâmetros para esses modelos analisando *corpora* bilíngues. Essa abordagem dispensa os linguistas: os programadores responsáveis pela construção desses sistemas não precisam nem mesmo falar a língua com a qual estão trabalhando.[68]

Nos últimos anos, programas de reconhecimento facial têm se aperfeiçoado suficientemente a ponto de serem utilizados na automatização do controle de fronteiras na Europa e na Austrália. O Departamento de Estado dos Estados Unidos opera um sistema de reconhecimento facial com mais de 75 milhões de fotografias para o processamento de vistos. Sistemas de vigilância utilizam IA cada vez mais sofisticada e tecnologias de mineração de dados para analisar voz, vídeos ou texto, sendo que boa parte desses dados são coletados em meios de comunicação eletrônicos no mundo todo e armazenados em enormes data centers.

Programas que provam teoremas e resolvem equações estão agora tão estabelecidos que dificilmente são reconhecidos como IA. Solucionadores de equações estão incluídos em programas de computação científica tal como o Mathematica. Métodos formais de verificação, incluindo provadores automáticos de teoremas, são rotineiramente utilizados por fabricantes de chips para verificar o comportamento dos projetos de circuitos antes de sua produção.

As organizações militares e de inteligência norte-americanas têm liderado a implementação em larga escala de robôs desarmadores de bombas, drones de vigilância e ataque e outros veículos não tripulados. Estes ainda dependem sobretudo do controle remoto operado por humanos, mas o trabalho para aumentar suas capacidades autônomas está em andamento.

O escalonamento inteligente é uma importante área de sucesso. A ferramenta DART para planejamento e escalonamento logístico automático foi usada na operação

Tempestade no Deserto em 1991, com tal efeito que a Agência de Projetos de Pesquisa Avançada de Defesa (DARPA, sigla em inglês) dos Estados Unidos afirma que essa única aplicação mais do que pagou seus trinta anos de investimentos em IA.[69] Sistemas de reserva de voos usam sofisticados sistemas de escalonamento e precificação. Empresas fazem amplo uso de técnicas de IA em sistemas de controle de estoque. Também são empregados sistemas automáticos de reserva por telefone e linhas de atendimento ao consumidor conectadas a softwares de reconhecimento de voz utilizados para conduzir seus clientes insatisfeitos através do labirinto de opções interdependentes do menu.

Tecnologias de IA estão por trás de muitos serviços na internet. Programas de computador policiam mundialmente o tráfego de e-mails e, apesar da contínua adaptação de *spammers* para driblar as contramedidas impostas a eles, filtros bayesianos *antispam* têm conseguido controlar o problema. Programas de computador que utilizam componentes de IA são responsáveis por aprovar ou reprovar automaticamente transações realizadas com cartões de crédito e monitoram continuamente a atividade das contas procurando sinais de uso fraudulento. Sistemas de recuperação de informação também fazem uso extensivo do aprendizado de máquina. A ferramenta de busca do Google é, provavelmente, o maior sistema de IA já criado.

Agora, é importante ressaltar que a demarcação entre inteligência artificial e programas de computador em geral não é tão nítida. Algumas das aplicações listadas acima podem ser vistas mais como aplicações genéricas de software do que especificamente como IA — embora isso nos traga de volta à máxima de McCarthy, de que, quando algo funciona, não é mais chamado de IA. Uma distinção mais relevante para nossos propósitos é aquela que existe entre sistemas que têm uma gama limitada de capacidade cognitiva (sejam eles chamados de “IA” ou não) e sistemas que possuem capacidade de solução de problemas de aplicação mais geral. Essencialmente todos os sistemas atualmente em uso são do primeiro tipo: limitado. Contudo, muitos deles possuem componentes que podem também ter participação em uma futura inteligência artificial geral ou estar a serviço de seu desenvolvimento — componentes como classificadores, algoritmos de busca, planejadores, solucionadores e estruturas representacionais.

Um ambiente de alto risco e extremamente competitivo no qual sistemas de IA operam atualmente é o mercado financeiro global. Sistemas de negociação de ações automatizados são amplamente usados por grandes corretoras de investimentos. Enquanto alguns deles são simplesmente utilizados para automatizar a execução de uma ordem de compra ou venda emitida por um administrador de fundos humano, outros seguem complicadas estratégias de compra e venda que se adaptam às mudanças das condições do mercado. Sistemas analíticos usam uma variedade de tecnologias de mineração de dados e análise de séries temporais para mapear padrões e tendências em mercados de valores mobiliários ou para correlacionar os movimentos históricos dos preços com variáveis externas tais como palavras-chave em painéis eletrônicos de notícias. Agências de notícias financeiras vendem *feeds* de notícias que são formatados especialmente para o uso de tais programas de IA. Outros sistemas se especializaram em encontrar

oportunidades de arbitragem entre mercados ou dentro deles, ou em negociações de alta frequência que buscam lucrar com pequenos movimentos dos preços que ocorrem em questão de milissegundos (uma escala de tempo na qual as latências de comunicação, mesmo para sinais na velocidade da luz em cabos de fibra ótica, se tornam significativos, tornando vantajosa a localização de computadores próximos ao local da transação). Negociações algorítmicas de alta frequência são responsáveis por mais da metade das transações envolvendo ações ordinárias comercializadas nos mercados norte-americanos.[70] A negociação algorítmica estava envolvida no Flash Crash de 2010 (ver quadro 2).

QUADRO 2

O Flash Crash de 2010

Na tarde do dia 6 de maio de 2010, os mercados de ações dos Estados Unidos já apresentavam uma queda de 4% em virtude da preocupação com a crise de endividamento europeia. Às 14h32, um grande vendedor (um conglomerado de fundos de investimento) iniciou um algoritmo de venda para liquidar um grande número de contratos futuros E-Mini S&P 500 a uma taxa de venda relacionada à liquidez medida minuto-a-minuto na transação. Esses contratos foram comprados por algoritmos de negociação de alta frequência que estavam programados para eliminar rapidamente suas posições longas temporárias vendendo os contratos para outros negociadores. Com uma fraca demanda por parte de compradores fundamentais, os algoritmos de negociação começaram a vender E-Minis principalmente para outros algoritmos de negociação, criando um efeito de "batata quente" que elevou o volume das negociações — isso foi interpretado pelo algoritmo de venda como um indicador de alta liquidez, o que induziu o aumento da taxa com a qual os contratos E-Mini eram colocados no mercado, elevando a tendência de queda. Num dado momento, os negociadores de alta frequência começaram a se retirar do mercado, secando a liquidez enquanto os preços continuavam caindo. Às 14h45, a comercialização de E-Minis foi interrompida pelo *circuit breaker*, um mecanismo automático de interrupção das transações da bolsa. Quando a comercialização foi reiniciada, apenas cinco segundos depois, os preços estabilizaram e logo começaram a recuperar boa parte de suas perdas. Mas, por um momento, no auge da crise, um trilhão de dólares desapareceu do mercado, e efeitos de propagação fizeram com que um número substancial de transações de títulos individuais fosse realizado a preços "absurdos", tais como 1 centavo ou 100 mil dólares. Após o fechamento do mercado naquele dia, os representantes das bolsas de valores se encontraram com as entidades reguladoras e decidiram cancelar todas as transações que haviam sido executadas com valores 60% (ou mais) maiores que os níveis anteriores à crise (considerando tais transações "claramente errôneas" e, por isso, sujeitas ao cancelamento *post facto*, conforme previsto nas leis de negociação de ações).[71]

Recontar aqui esse episódio é uma digressão porque os programas de computação envolvidos no Flash Crash não eram particularmente inteligentes ou sofisticados e o tipo de ameaça que eles criaram é fundamentalmente diferente das preocupações que levantaremos a seguir em relação à perspectiva da superinteligência de máquina. Entretanto, esses eventos ilustram uma série de lições úteis. Uma delas é o alerta de que interações entre componentes individualmente simples (como o algoritmo de venda e os algoritmos de negociação de alta frequência) podem produzir resultados complexos e inesperados. Riscos sistêmicos podem se acumular em um sistema conforme novos elementos são introduzidos, riscos que não são óbvios até o momento em que algo dá errado (e algumas vezes nem mesmo quando isso acontece).[72]

Outra lição é a de que profissionais inteligentes podem fornecer instruções para um programa com base em uma suposição que pareça sensata e razoável (por exemplo, que o volume de transações é uma boa medida da liquidez do mercado), e isso pode produzir resultados catastróficos quando o programa continua a agir rigorosamente dentro da lógica segundo a qual foi instruído, mesmo em situações não previstas na qual tal suposição se mostra inválida. O algoritmo só faz o que ele está programado para fazer e, a menos que ele seja um tipo muito especial de algoritmo, ele não dá a mínima para o fato de ficarmos horrorizados com suas ações totalmente inapropriadas. Esse é um tema do qual falaremos mais adiante.

A terceira observação em relação ao Flash Crash é que, embora a automação tenha contribuído para o incidente, ela contribuiu também para a sua solução. A lógica de interrupção automática pré-programada, que suspendeu a comercialização quando os preços saíram do controle, foi programada para entrar automaticamente em ação, pois se antecipou corretamente que eventos passíveis de acionar a interrupção poderiam acontecer numa escala de tempo muito curta para a capacidade de resposta dos humanos. A necessidade de funcionalidades de segurança pré-instaladas e de acionamento automático — em oposição à

confiança no tempo de resposta de uma supervisão humana — mais uma vez antecipa um tema que será importante em nossa discussão sobre a superinteligência de máquina.[73]

Opiniões sobre o futuro da inteligência de máquina

O progresso em duas grandes frentes — por um lado, em direção a uma base mais sólida em estatística e teoria da informação para aprendizagem de máquina, e, por outro, em direção ao sucesso prático e comercial de várias aplicações de problemas ou áreas específicas — tem trazido de volta à pesquisa de IA parte do prestígio que havia sido perdido. Pode haver, no entanto, um efeito cultural residual na comunidade de IA, proveniente da fase inicial da sua história, que faz com que muitos dos principais pesquisadores relutem em se alinhar com ambições exageradas. Por causa disso, Nils Nilsson, um dos veteranos na área, se queixa de que seus colegas atuais não possuem a ousadia de espírito que impulsionou os pioneiros de sua própria geração:

> Acredito que a preocupação por "respeitabilidade" tem tido um efeito estupidificante sobre alguns pesquisadores de IA. Eu os ouço dizer coisas como: "A IA costumava ser criticada por sua superficialidade. Agora que temos feito progresso sólido, não devemos arriscar nossa respeitabilidade". As consequências desse conservadorismo têm sido uma concentração maior na "IA fraca" — a variedade dedicada a prover ajuda ao pensamento humano — e o afastamento de uma "IA forte" — a variedade que busca mecanizar a inteligência de nível humano.[74]

O sentimento de Nilsson foi ecoado por muitos outros fundadores, incluindo Marvin Minsky, John McCarthy e Patrick Winston.[75]

Os últimos anos têm assistido a um ressurgimento do interesse em IA, o que pode ainda conduzir a esforços renovados em direção à inteligência artificial *geral* (o que Nilsson chama de "IA forte"). Além de um hardware mais rápido, um projeto contemporâneo se beneficiaria dos grandes avanços que têm sido alcançados nas muitas subáreas de IA, em engenharia de software de forma geral e nas áreas correlatas, como a neurociência computacional. Um indício de que existe uma demanda reprimida por informação de qualidade e cursos de formação pode ser percebido através da resposta à oferta de um curso on-line gratuito de introdução à inteligência artificial na Universidade Stanford no outono de 2011, organizado por Sebastian Thrun e Peter Norvig. Aproximadamente 160 mil estudantes de várias partes do mundo se inscreveram para participar (e 23 mil concluíram o curso).[76]

A opinião de especialistas sobre o futuro da IA varia consideravelmente. Há tanto divergências sobre a escala de tempo quanto sobre as formas que a IA pode eventualmente assumir. Um estudo recente notou que previsões sobre o desenvolvimento futuro da inteligência artificial "são tão confiantes quanto diversas".[77]

Embora a distribuição das opiniões atuais não tenha sido cuidadosamente medida, podemos ter uma vaga ideia a partir de pesquisas menores e observações

informais. Em particular, uma série de pesquisas recentes têm contado com entrevistas a membros de vários grupos relevantes de especialistas, que são questionados a respeito de quando, segundo eles, a "inteligência de máquina de nível humano" (HLMI, na sigla em inglês) poderia ser desenvolvida, sendo esta definida como "aquela que pode executar a maioria das profissões humanas ao menos tão bem quanto um humano típico".[78] Os resultados são mostrados na tabela 2. A amostra combinada resultou na estimativa (mediana) a seguir: 10% de probabilidade de que a HLMI seja alcançada até 2022, 50% de probabilidade até 2040 e 90% de probabilidade até 2075. (Foi solicitado aos entrevistados que condicionassem suas estimativas à premissa de que a "atividade científica humana prossiga sem maiores perturbações negativas".)

Esses números devem ser encarados com certa reserva: as amostras são pequenas e não necessariamente representam a população geral de especialistas. Eles estão, no entanto, de acordo com resultados de outras pesquisas.[79]

Os resultados das pesquisas estão também em conformidade com entrevistas realizadas com aproximadamente duas dezenas de pesquisadores de campos relacionados à IA e recentemente publicadas. Por exemplo, Nils Nilsson passou uma longa e produtiva carreira trabalhando em problemas de busca, planejamento, representação de conhecimento e robótica; escreveu livros acadêmicos a respeito da inteligência artificial; e completou recentemente o trabalho mais abrangente sobre a história da IA escrito até hoje.[80] Quando questionado sobre o advento da HLMI, ele opinou da seguinte forma:[81]

10% de chance: 2030 **50%** de chance: 2050 **90%** de chance: 2100

TABELA 2

Quando a inteligência de máquina de nível humano será alcançada?*

	10%	**50%**	**90%**
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Combinados	2022	2040	2075

* A tabela mostra os resultados de quatro pesquisas de opinião diferentes, bem como os resultados combinados. As duas primeiras foram pesquisas realizadas em conferências acadêmicas: PT-AI, participantes da conferência *Philosophy and Theory of AI*, em Salônica, no ano de 2011 (os participantes foram entrevistados em novembro de 2012), com uma taxa de resposta de 43 num conjunto de 88; e AGI, participantes das

conferências *Artificial General Intelligence e Impacts and Risks of Artificial General Intelligence*, ambas realizadas em Oxford em dezembro de 2012 (taxa de resposta: 72/111). A pesquisa EETN é uma amostragem dos membros da Greek Association for Artificial Inteligence, uma organização profissional de pesquisadores da área, em abril de 2013 (taxa de resposta: 26/250). A pesquisa TOP100 coletou as opiniões dos 100 principais autores em inteligência artificial de acordo com um índice de citação, em maio de 2013 (grau de resposta: 29/100).

A julgar pelas transcrições publicadas das entrevistas, a distribuição de probabilidade do professor Nilsson se mostra representativa da opinião de muitos especialistas da área — embora, novamente, deve ser enfatizado que há uma ampla gama de opiniões: há profissionais que são consideravelmente mais confiantes no progresso rápido, esperando a HLMI entre os anos 2020 e 2040, e outros que estão confiantes de que ela nunca será alcançada ou que esteja indefinidamente distante.[82] Além disso, alguns entrevistados sentem que a noção de um "nível humano" de inteligência artificial é mal definida ou enganosa, ou os entrevistados se mostram, por outras razões, relutantes em deixar registrada uma previsão quantitativa.

Minha opinião é que os números medianos relatados na pesquisa com especialistas não possuem massa de probabilidade suficiente nas datas mais avançadas. Uma probabilidade de 10% de que a HLMI não esteja desenvolvida em 2075 ou mesmo em 2100 (com a condição de que a "atividade científica humana prossiga sem grandes perturbações negativas") me parece muito baixa.

Historicamente, os pesquisadores de IA não têm se mostrado muito bons em prever o ritmo dos avanços em seu próprio campo ou a forma que tais avanços poderiam assumir. Por um lado, algumas tarefas, como jogar xadrez, se tornaram possíveis através de programas surpreendentemente simples; e céticos que diziam que as máquinas "nunca" seriam capazes de fazer isso ou aquilo se mostraram equivocados. Por outro lado, os erros mais comuns entre os profissionais têm sido subestimar as dificuldades de fazer com que um sistema execute, de forma consistente, tarefas do mundo real, e o de superestimar as vantagens de seus próprios projetos ou técnicas preferidas.

A pesquisa também incluiu duas outras perguntas relevantes para a nossa investigação. Uma delas questionou os entrevistados a respeito do tempo que, segundo eles, seria necessário para alcançar a superinteligência, supondo que a inteligência de máquina com nível humano seja atingida. Os resultados estão na tabela 3.

Outra pergunta levantou a questão sobre o tipo de impacto geral de longo prazo que poderia ser causado na humanidade com o alcance da inteligência de máquina de nível humano. As respostas estão resumidas na figura 2.

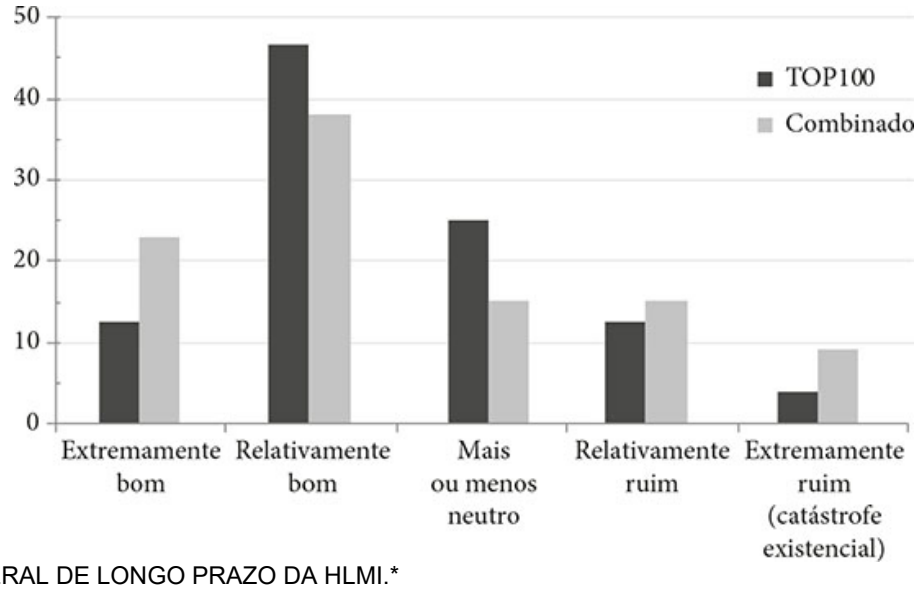
TABELA 3

Quanto tempo será necessário para que o nível humano alcance a superinteligência?

	Dentro de 2 anos após HLMI	**Dentro de 30 anos após HLMI**

TOP100	5%	50%
Combinado	10%	75%

FIGURA 2



IMPACTO GERAL DE LONGO PRAZO DA HLMI.*
**O diagrama mostra estimativas usando medianas renormalizadas. As médias são significativamente diferentes. Por exemplo, as estimativas médias para um desfecho "extremamente ruim" foram de 7,6% (para o TOP100) e 17,2% (para o conjunto total de especialistas entrevistados).*

Novamente meu entendimento difere, de alguma forma, das opiniões expressas na pesquisa. Atribuo uma probabilidade maior de que a superinteligência seja criada num ritmo relativamente rápido após o alcance da inteligência de máquina de nível humano. Também tenho uma visão mais polarizada sobre as consequências, acreditando que um resultado extremamente bom ou extremamente ruim é mais provável que um resultado equilibrado. As razões para tanto serão melhor esclarecidas no decorrer do livro.

Amostras pequenas, vieses de seleção e — acima de tudo — a baixa confiabilidade inerente às opiniões subjetivas solicitadas indicam que não se deve tirar muitas conclusões a respeito dessas pesquisas de opinião e entrevistas com especialistas. Elas não nos permitem extrair qualquer conclusão sólida, mas fornecem uma conclusão ampla. Elas sugerem que (ao menos na ausência de melhores dados ou análises) pode ser razoável acreditar que a inteligência de máquina de nível humano tem boas chances de ser desenvolvida até a metade do século e que há uma probabilidade considerável de que seja desenvolvida bem mais cedo ou muito mais tarde; que ela possa talvez, logo depois, resultar na superinteligência; e que uma ampla gama de consequências poderá ocorrer, incluindo consequências extremamente boas e consequências tão ruins quanto a extinção da espécie humana.[83] No mínimo, elas sugerem que é válido analisar o tema atentamente.

02

CAPÍTULO 2

CAMINHOS PARA A SUPERINTELIGÊNCIA

As máquinas são, atualmente, muito inferiores aos humanos em inteligência geral. Entretanto, um dia (como já sugerimos), elas serão superinteligentes. Como faremos para, a partir do estágio atual, atingir o estágio da superinteligência das máquinas? Este capítulo explora diversos caminhos tecnológicos possíveis. Falaremos a respeito de inteligência artificial, emuladores de atividade cerebral, cognição biológica, interfaces homem-máquina e também sobre redes e organizações. Avaliaremos seus diferentes graus de plausibilidade como caminhos para a superinteligência. A existência de múltiplos caminhos aumenta a probabilidade de que a superinteligência seja alcançada pelo menos por um desses caminhos.

Podemos provisoriamente definir uma superinteligência como *qualquer intelecto que exceda em muito o desempenho cognitivo dos seres humanos em, virtualmente, todos os domínios de interesse*.[84] Teremos mais a dizer sobre o conceito de superinteligência no próximo capítulo, quando esse conceito será submetido a uma espécie de análise espectral, buscando distinguir algumas possíveis formas de superinteligência. Mas, por ora, a definição aproximada aqui exposta será suficiente. Perceba que essa definição não trata a respeito da forma como a superinteligência será implementada. Essa definição tampouco tem qualquer preocupação com os *qualia*: o fato de uma superinteligência alcançar uma experiência subjetiva consciente pode ser de grande importância para a discussão de algumas questões (em particular, aquelas de cunho moral), mas aqui o foco principal está voltado para os antecedentes causais e as consequências da superinteligência, e não para a metafísica da mente.[85]

De acordo com essa definição, o programa de xadrez Deep Fritz não é uma superinteligência, uma vez que Fritz é inteligente apenas no que diz respeito aos jogos de xadrez. Alguns tipos de superinteligência de domínio específico podem, entretanto, ser importantes. Quando nos referirmos ao desempenho de uma superinteligência que se encontra restrita a um domínio particular, essa limitação será explicitamente observada. Por exemplo, uma “superinteligência da engenharia” seria um intelecto que superasse em muito as capacidades humanas na área de engenharia. Caso não seja explicitado de outra maneira, usaremos o termo superinteligência em referência apenas a sistemas que tenham um nível sobre-humano de inteligência *geral*.

Mas como poderíamos criar uma superinteligência? Examinemos alguns caminhos possíveis.

Inteligência artificial

Os leitores deste capítulo não devem esperar encontrar instruções para a programação de uma inteligência artificial geral. Essas instruções, é claro, ainda não existem. E caso eu as possuísse, certamente não iria publicá-las num livro. (Se as razões para tal decisão não são imediatamente óbvias, os argumentos nos capítulos subsequentes serão capazes de esclarecê-las.)

Podemos, entretanto, descrever alguns elementos gerais que precisariam estar presentes em um sistema dessa natureza. A esta altura, parece claro que a capacidade de aprendizagem seria uma das características centrais de um sistema que tivesse como intuito atingir a inteligência geral, e não algo a ser acrescentado posteriormente como uma extensão ou um adendo. O mesmo vale para a habilidade de lidar efetivamente com incertezas e informações probabilísticas. Entre outras características que provavelmente deverão ser centrais em um sistema moderno de IA que objetiva atingir a inteligência geral, podemos listar a capacidade de elaboração de conceitos úteis a partir da análise de dados sensoriais externos e estados internos ao agente e a habilidade de aplicar esses conceitos a uma série de representações combinatórias flexíveis que possam ser usadas na elaboração de raciocínios lógicos e intuitivos.

Na maioria dos casos, os primeiros sistemas GOFAI não focaram na capacidade de aprendizagem, em incertezas ou na formação de conceitos, muito provavelmente porque as técnicas utilizadas para lidar com esses aspectos ainda não estavam plenamente desenvolvidas. Isso não quer dizer que essas ideias sejam muito recentes. A noção de que a capacidade de aprendizagem seria capaz de fazer com que um sistema mais simples atingisse níveis humanos de inteligência data, pelo menos, de 1950, quando Alan Turing descreveu o conceito de “máquina criança”:

> Em vez de tentar criar um programa capaz de simular a mente de um adulto, por que não tentar produzir um que simule a mente de uma criança? Se ele fosse, então, submetido a uma trajetória apropriada de aprendizado, seríamos capazes de obter o cérebro de um adulto.[86]

Turing contemplou um processo iterativo para o desenvolvimento da “máquina criança”:

> Não podemos esperar encontrar uma boa máquina criança logo na primeira tentativa. É possível fazer uma experiência, ensinando uma dessas máquinas para verificar a qualidade de seu processo de aprendizagem. Podemos, então, tentar com uma outra máquina e checar se ela se sai melhor ou pior. Existe uma conexão óbvia entre esse processo e a evolução [...]. Pode-se esperar, entretanto, que esse processo seja mais eficiente do que a evolução. A sobrevivência do mais apto é um processo lento demais para medir vantagens. O cientista, ao exercitar sua inteligência, deveria ser capaz de acelerá-lo. Igualmente importante é o fato de tal processo não estar restrito a mutações aleatórias. Se o cientista conseguir perceber a causa de alguma deficiência, ele

> provavelmente será capaz de pensar em algum tipo de mutação capaz de corrigi-la.[87]

Sabemos que processos evolutivos aleatórios são capazes de produzir níveis humanos de inteligência geral, levando em conta que isso já aconteceu pelo menos uma vez. Sendo assim, processos evolutivos planejados — isto é, programas genéticos elaborados e executados por um programador humano inteligente — poderiam ser capazes de alcançar resultados similares com muito mais eficiência. Essa mesma observação já foi usada por outros filósofos e cientistas, incluindo David Chalmers e Hans Moravec, para argumentar que a IA em níveis humanos não apenas é teoricamente possível como bastante plausível de ser alcançada ainda neste século.[88] A ideia é que podemos estimar as potencialidades relativas da evolução natural e da engenharia humana para produzir inteligência e, ao fazer isso, perceberemos que a engenharia humana já é, em algumas áreas, muito superior à evolução natural e provavelmente não demorará muito para que o seja em todas as demais áreas. Nessa perspectiva, o fato de que a evolução natural tenha sido capaz de produzir inteligência indica que a engenharia humana será, muito em breve, capaz de fazer o mesmo. A esse respeito, Moravec escreveu (já em 1976):

> A existência de diversos exemplos de inteligência projetadas dentro dessas limitações nos dá bastante confiança de que seremos capazes de fazer o mesmo muito em breve. A situação é análoga ao caso do voo mais pesado que o ar, em que pássaros, morcegos e insetos comprovaram claramente essa possibilidade, muito antes de nossa cultura tê-la dominado.[89]

É preciso, entretanto, ter cautela sobre o que se pode inferir a partir dessa linha de raciocínio. É verdade que a evolução natural produziu o voo mais pesado que o ar e que a engenharia humana conseguiu posteriormente fazer o mesmo (embora tenha utilizado um mecanismo completamente diferente). Podemos citar outros exemplos nessa mesma linha, como o sonar, a navegação por magnetismo, armas químicas, fotorreceptores e vários tipos de características de desempenho mecânico e cinético. Entretanto, é igualmente possível apontar áreas nas quais os engenheiros humanos não foram capazes até agora de se equiparar à evolução natural: morfogênese, autorregeneração e defesa imunológica são alguns exemplos de áreas em que os esforços humanos não chegaram nem próximos ao que a natureza foi capaz de desenvolver. Sendo assim, o argumento de Moravec não é capaz de nos dar “bastante confiança” de que seremos capazes de desenvolver a IA em níveis humanos “muito em breve”. Na melhor das hipóteses, a evolução da vida inteligente coloca um limite superior para a dificuldade intrínseca de se criar inteligência. Mas esse limite superior pode estar bem acima das capacidades atuais da engenharia humana.

Outra maneira de aplicar um argumento evolutivo na questão da viabilidade da IA é através da ideia de que, ao rodar algoritmos genéticos em computadores suficientemente rápidos, seríamos capazes de alcançar resultados comparáveis

àqueles da evolução biológica. Essa versão do argumento evolutivo propõe, por conseguinte, um método específico através do qual a inteligência poderia ser gerada.

Mas seria verdade que em breve conseguiremos desenvolver capacidade computacional suficiente para replicar os processos evolutivos relevantes que culminaram na inteligência humana? A resposta para essa pergunta depende do quanto a tecnologia computacional se desenvolverá nas próximas décadas e do poder de processamento necessário para a execução de algoritmos genéticos de forma tão otimizada quanto o processo evolutivo de seleção natural que nos precede. No fim das contas, embora a conclusão a que chegamos ao desenvolver essa linha de raciocínio decepcione por seu grau de indeterminação, é instrutiva a tentativa de se fazer uma estimativa, mesmo que imprecisa (ver quadro 3). Pelo menos, esse exercício chama a atenção para alguns pontos interessantes que ainda nos são desconhecidos.

O fato é que os recursos computacionais necessários para que sejamos capazes de simplesmente replicar os processos evolutivos relevantes que produziram a inteligência de nível humano estão ainda muito longe de nosso alcance — e permanecerão assim, mesmo que a lei de Moore continuasse por mais um século (ver figura 3). É plausível, entretanto, que, em comparação a uma replicação força-bruta dos processos evolutivos naturais, grandes ganhos de eficiência possam ser alcançados com a projeção de um processo de busca que tenha a inteligência como *objetivo*, usando várias melhorias óbvias em relação à seleção natural. Mesmo assim é muito difícil prever a magnitude dos ganhos de eficiência de tal processo evolutivo artificial. Não somos sequer capazes de afirmar se o ganho seria de cinco ou 25 ordens de magnitude. Sendo assim, na ausência de mais informações, argumentos evolutivos não são capazes de restringir significativamente nossas expectativas nem em relação às dificuldades referentes à criação de uma máquina com inteligência de nível humano, nem quanto ao tempo que seria necessário para o seu desenvolvimento.

QUADRO 3

O que seria necessário para replicar a evolução?

Nem todos os feitos alcançados pelo processo evolutivo durante o desenvolvimento da inteligência humana são relevantes a um engenheiro interessado em fazer evoluir artificialmente uma máquina inteligente. Apenas uma pequena parte da seleção evolutiva em nosso planeta envolveu a seleção por inteligência. Mais especificamente, alguns obstáculos difíceis de serem superados pelos engenheiros humanos podem ter sido alvo de apenas uma pequena porção do processo de seleção evolutiva. Por exemplo, como nossos computadores funcionam com energia elétrica, não é necessário que tenhamos que reinventar as moléculas responsáveis pela economia de energia celular para desenvolver máquinas inteligentes — entretanto, boa parte da seleção natural observada no processo evolutivo da Terra pode ter sido usada para a evolução molecular dos caminhos metabólicos.[90]

É possível argumentar que a chave para o desenvolvimento da IA pode ser encontrada na estrutura dos sistemas nervosos que surgiram a menos de 1 bilhão de anos atrás.[91] Se seguirmos por essa linha, o número de “experimentos” relevantes ao desenvolvimento de um processo evolutivo artificial diminui consideravelmente. Há algo em torno de $4–6\times10^{30}$ procariontes no mundo hoje em dia, mas apenas 10^{19} insetos e menos de 10^{10} humanos (sendo que a população humana antes do desenvolvimento da agricultura era muito menor).[92] Esses números são apenas moderadamente intimidadores.

Todavia, algoritmos evolutivos não precisam apenas de variação entre indivíduos para fazer a seleção, sendo também necessário que eles possuam uma função de aptidão para avaliar indivíduos, e esse é, geralmente, o elemento computacional mais dispendioso. É plausível imaginar que, no que se refere à evolução de uma inteligência artificial, uma função de aptidão necessite simular o desenvolvimento neural, a capacidade de aprendizagem e a cognição de um indivíduo para que consiga, então, avaliar o seu grau de aptidão. Desse modo, talvez seja melhor não olhar para o número absoluto de organismos com sistemas nervosos complexos, mas, em vez disso, estar atento à quantidade de neurônios presentes nos organismos biológicos que talvez seja necessário simular para imitar a função de aptidão do processo evolutivo natural. É possível fazer uma estimativa simples desses números levando em consideração os insetos que dominam a biomassa animal terrestre (apenas as formigas contribuem com algo em torno de 15% a 20% da biomassa).[93] O tamanho do cérebro dos insetos varia consideravelmente, sendo que os insetos de maior porte e os insetos sociais apresentam os maiores cérebros: o cérebro de uma abelha tem pouco menos de 10^6 neurônios, o de uma mosca-das-frutas possui 10^5 neurônios e o das formigas se encontra no meio do caminho, com 250 mil neurônios.[94] A grande maioria dos insetos menores possui cérebros com apenas alguns milhares de neurônios. Levando em conta valores conservadoramente altos, se estipularmos que todos os 10^{19} insetos possuem números de neurônios iguais aos de uma mosca-das-frutas, o valor total do número de neurônios de insetos no mundo seria de 10^{24}. Esse valor poderia ainda ser aumentado em uma ordem de magnitude para englobar também os microrganismos aquáticos, aves, répteis, mamíferos etc., chegando ao número de 10^{25}. (Comparativamente, durante o período anterior ao desenvolvimento da agricultura, havia menos de 10^7 seres humanos, cada um com menos de 10^{11} neurônios: sendo assim, havia menos de 10^{18} neurônios humanos ao todo, embora os seres humanos apresentem uma quantidade superior de sinapses por neurônio.)

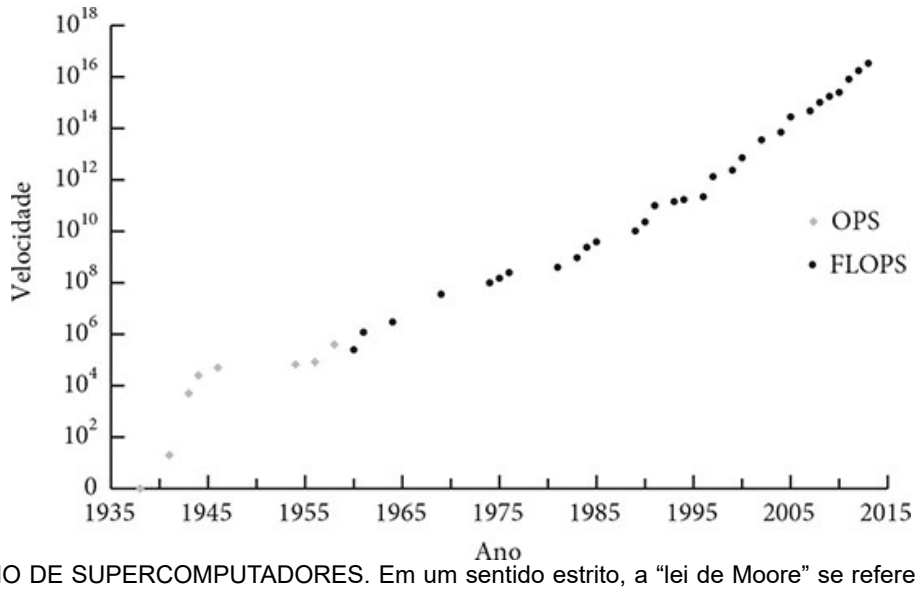
O custo computacional necessário para simular um único neurônio depende do nível de detalhe que se pretende incluir na simulação. Modelos neurais extremamente simples utilizam em torno de mil operações de ponto flutuante por segundo (FLOPS, na sigla em inglês) para simular um neurônio (em tempo real). O modelo eletrofisiologicamente realista de Hodgkin-Huxley utiliza 1.200.000 FLOPS. Um modelo multicompartimentado mais detalhado acrescentaria outras três ou quatro ordens de magnitude, enquanto modelos mais abstratos, que representam diretamente sistemas de neurônios, poderiam subtrair duas ou três ordens de magnitude dos modelos mais simples.[95] Se fôssemos simular 10^{25} neurônios por um período de 1 bilhão de anos de evolução (um período mais longo do que o da existência de sistemas nervosos como os conhecemos), e deixássemos nossos computadores executando essa simulação por um ano, seria necessária uma capacidade computacional da ordem de 10^{31} a 10^{44} FLOPS. Comparativamente, o supercomputador chinês Tianhe-2, considerado o mais poderoso do mundo em setembro de 2013, é capaz de prover apenas $3{,}39\times10^{16}$ FLOPS. Nas últimas décadas, os computadores comerciais têm levado aproximadamente 6,7 anos para aumentar sua capacidade de processamento em uma ordem de magnitude. Mesmo um século de contínua aplicação da lei de Moore não seria capaz de preencher essa lacuna. Utilizar hardware mais especializado ou permitir períodos mais longos de execução contribuiria apenas para reduzir a lacuna em algumas poucas ordens de magnitude.

Esses números também são conservadores em outro aspecto. A evolução produziu inteligência humana sem ter sido essa a sua intenção. Dito de outro modo, a função de aptidão dos organismos naturais não realiza a seleção natural apenas em busca da inteligência e de seus precursores.[96] Mesmo em ambientes nos quais organismos com capacidade superior de processamento de informação possam obter diversos benefícios, a seleção por inteligência não é garantida, pois melhorias que visam o desenvolvimento da inteligência podem impor custos significativos (e isso, em geral, acontece), como um consumo maior de energia ou um período maior de amadurecimento, e esses custos podem não compensar os ganhos adquiridos por um comportamento mais inteligente. Ambientes extremamente agressivos também reduzem o valor da inteligência: quanto mais curta a expectativa de vida, menor é o tempo disponível para que uma maior capacidade de aprendizagem gere frutos. Uma redução na pressão seletiva por inteligência desacelera a disseminação de inovações capazes de aprimorar a inteligência, desacelerando também as oportunidades para que a seleção natural favoreça inovações subsequentes que dependam das anteriores. Além disso, a evolução pode ficar presa em ótimos locais, o que poderia ser percebido pelos humanos e contornado por meio da alteração do equilíbrio entre exploração e aproveitamento ou com o fornecimento de uma progressão suave de testes de inteligência cada vez mais complicados.[97] Como já foi mencionado, a evolução dispersa muito de seu poder de seleção em virtude de fatores que não possuem nenhuma relação com a inteligência (por exemplo, nas “corridas da rainha Vermelha” de coevolução competitiva entre sistemas imunológicos e parasitas). A evolução continua a gastar recursos produzindo mutações que provaram ser consistentemente letais e não é eficiente em tirar vantagem de similaridades estatísticas entre os efeitos relativos a diferentes mutações. Essas são deficiências da seleção natural (quando vista como um meio de se criar inteligência) que seriam relativamente simples de ser evitadas pelos engenheiros com o uso de algoritmos evolutivos para o desenvolvimento de software inteligente.

Provavelmente, depois de eliminadas ineficiências como as descritas acima, seria possível diminuir em muitas ordens de magnitude nosso cálculo anterior de 10^{31}–10^{44} FLOPS. Infelizmente, é difícil saber ao certo de quantas ordens de magnitude estamos falando. É difícil até mesmo fazer uma estimativa — até onde sabemos, os valores poupados com essa abordagem mais eficiente poderiam ser de cinco ordens de magnitude, ou dez, ou 25.[98]

Existe ainda mais um complicador em relação a esse tipo de consideração evolucionária que torna ainda mais difícil estimar o limite superior do grau de dificuldade da criação de um processo evolutivo que objetive a inteligência. Devemos evitar o erro de inferir que, como a inteligência evoluiu na Terra, o processo evolucionário por trás disso possuía uma alta probabilidade a *priori* de produzir inteligência. Tal suposição não se sustenta, pois não leva em consideração o efeito de observação seletiva, que garante que todos os observadores são originários de um planeta no qual a vida inteligente surgiu, não importando o quão provável, ou improvável, seria a possibilidade de qualquer planeta produzir inteligência. Suponha, por exemplo, que além dos efeitos sistemáticos da seleção natural, também fosse necessária uma grande dose de *sorte* e *coincidência* para a produção de vida inteligente — uma dose tão grande que a vida inteligente se desenvolvesse apenas em um de 10^{30} planetas nos quais replicadores simples existissem. Nesse caso, quando rodássemos nossos algoritmos genéticos na tentativa de reproduzir o que a seleção natural fez, logo descobriríamos que seria necessário rodar algo em torno de 10^{30} simulações antes que fôssemos capazes de encontrar um modelo no qual todos os elementos se alinhassem exatamente da forma correta. Isso parece estar em conformidade com nossas observações de que a vida se desenvolveu aqui na Terra. Apenas através de considerações cuidadosas e complexas — analisando exemplos que mostram a evolução convergente de características ligadas à inteligência e percebendo sutilezas da teoria de observação seletiva — é que seremos capazes de ultrapassar, mesmo que parcialmente, essa barreira epistemológica. A não ser que alguém se dê ao trabalho de fazer isso, será impossível descartar a possibilidade de que o suposto “limite superior” do esforço computacional necessário para simular a evolução natural da inteligência, descrito no quadro 3, de fato seja trinta ordens de magnitude mais baixo do que deveria ser (ou qualquer outro valor significativo).[99]

FIGURA 3



DESEMPENHO DE SUPERCOMPUTADORES. Em um sentido estrito, a "lei de Moore" se refere à observação de que o número de transistores em um circuito integrado, já há muitas décadas, vem dobrando aproximadamente de dois em dois anos. Todavia, o termo tem sido usado com frequência para fazer referência a uma observação mais geral segundo a qual muitas métricas de desempenho em tecnologia computacional seguiram uma tendência semelhante de crescimento exponencial. Aqui, mostramos a velocidade máxima do supercomputador mais rápido do mundo como função de tempo (em uma escala vertical logarítmica). Nos últimos anos, o aumento na velocidade serial dos processadores se manteve estagnado, mas o crescente uso da paralelização tem possibilitado que o número total de computações realizadas continue seguindo a mesma tendência.*
**Wikipédia (2012b).*

Outra maneira de argumentar a favor da viabilidade da inteligência artificial é apontando para o cérebro humano e sugerindo que talvez pudéssemos usá-lo como um modelo para a inteligência de máquina. É possível distinguir algumas versões diferentes para essa abordagem, com base no grau de similaridade que o modelo pretenda ter com as funções cerebrais biológicas. Em um extremo — aquele que se encontra mais próximo à imitação — se encontra a ideia de *emulação completa do cérebro*, que discutiremos na próxima subseção. No outro extremo estão as abordagens inspiradas no funcionamento do cérebro, mas que não se propõem a uma imitação minuciosa. Avanços na neurociência e na psicologia cognitiva — que serão auxiliados, ainda, pelas melhorias nas técnicas de instrumentação — serão capaz de, eventualmente, desvendar os princípios gerais do funcionamento do cérebro. Esse conhecimento poderia, então, guiar os esforços em IA. Já falamos sobre redes neurais como um exemplo de técnica de IA inspirada no funcionamento do cérebro. A organização hierárquica da percepção é outra ideia que foi importada da neurociência para ser aplicada no aprendizado de máquina. Os estudos sobre a aprendizagem por reforço têm sido motivados (pelo menos em parte) por seu papel nas teorias psicológicas de cognição animal, e técnicas de aprendizagem por reforço (por exemplo, "algoritmos TD") inspiradas nessas teorias são, hoje em dia, amplamente utilizadas em IA.[100] Mais casos como esse certamente ocorrerão no futuro. Como existe um número limitado — talvez muito reduzido — de mecanismos fundamentais distintos que operam no cérebro, avanços incrementais contínuos na neurociência devem, em breve, descobrir todos eles. Entretanto, antes que isso

ocorra, é possível que uma abordagem híbrida, que combine técnicas inspiradas no funcionamento do cérebro com métodos puramente artificiais, também surta efeitos. Nesse caso, o sistema resultante não será necessariamente similar ao cérebro, ainda que mecanismos de funcionamento do cérebro tenham sido usados no seu desenvolvimento.

A disponibilidade do cérebro como um modelo fornece um forte ponto de sustentação para o argumento que prega a viabilidade da inteligência de máquina. Isso, entretanto, não nos torna capazes de prever quando ela será alcançada, pois é difícil estimar a velocidade com que se darão as descobertas futuras da neurociência. O que podemos afirmar é que quanto mais olhamos para o futuro, maior é a possibilidade de que os segredos da funcionalidade do cérebro tenham sido suficientemente decodificados a ponto de tornar possível a criação de uma inteligência de máquina dessa forma.

Diferentes pessoas trabalhando no desenvolvimento da inteligência de máquina possuem diferentes visões a respeito do quão promissoras as abordagens neuromórficas se mostram quando comparadas a abordagens totalmente sintéticas. A existência dos pássaros demonstrou que o voo mais pesado do que o ar era fisicamente possível e desencadeou esforços para a construção de máquinas voadoras. Entretanto, o primeiro avião funcional não batia asas. Só o tempo dirá se a inteligência de máquina será como o avião, que foi desenvolvido através de mecanismos artificiais, ou como a combustão, técnica que dominamos com a cópia da ocorrência natural do fogo.

A ideia de Turing de desenvolvimento de um software que adquira grande parte de seu conteúdo através da aprendizagem, em vez do conteúdo ser previamente programado, pode ser aplicada tanto numa abordagem neuromórfica como numa abordagem sintética para a inteligência de máquina.

Uma variação da concepção de Turing de uma máquina criança é a ideia de uma "IA embrionária" [em inglês, *seed* AI].[101] Já que uma máquina criança, no sentido em que foi concebida por Turing, teria uma arquitetura relativamente fixa que desenvolveria suas potencialidades apenas através do acúmulo prévio de conteúdo, uma IA embrionária seria uma forma mais sofisticada de inteligência artificial, capaz de aprimorar sua própria *arquitetura*. Nos estágios iniciais de uma IA embrionária, tais aprimoramentos poderiam ocorrer principalmente por tentativa e erro, aquisição de informação, ou com o auxílio de programadores. Já em estágios mais avançados, uma IA embrionária poderia ser capaz de *compreender* suficientemente bem seu próprio funcionamento a ponto de desenvolver, por conta própria, novos algoritmos e estruturas computacionais que aprimorassem seu desempenho cognitivo. Esse conhecimento poderia ser alcançado a partir do momento em que a IA embrionária alcançasse um nível suficiente de inteligência geral em diversos domínios do conhecimento, ou quando ela fosse capaz de cruzar uma barreira de conhecimento em algum campo particularmente relevante, como ciência da computação ou matemática.

Isso nos leva a outro conceito importante: o de "automelhoria recursiva". Uma IA embrionária que viesse a ser bem-sucedida seria capaz de aprimorar a si mesma iterativamente: uma primeira versão da IA poderia projetar uma versão melhor de si

mesma, essa segunda versão seria, por sua vez, mais inteligente que a primeira e igualmente capaz de produzir uma versão melhorada de si mesma e assim por diante.[102] Sob algumas condições, tal processo de automelhoria recursiva poderia continuar por tempo suficiente até resultar em uma explosão de inteligência — um evento no qual, em um curto espaço de tempo, o nível de inteligência de um sistema passaria de capacidades relativamente modestas de cognição (talvez sub-humanas em muitos aspectos, mas com talento específico para programar e realizar pesquisas em IA) a uma superinteligência radical. Retornaremos a essa importante possibilidade no capítulo 4, no qual a dinâmica de tal evento será analisada em maiores detalhes. Note que esse modelo sugere a possibilidade de algumas surpresas: tentativas de projetar uma inteligência artificial geral podem falhar completamente até que o último componente crítico seja adicionado, ponto a partir do qual uma IA embrionária poderia ser capaz de manter um padrão de automelhoria recursiva.

Antes de finalizarmos esta subseção, há mais um ponto que deve ser enfatizado: uma inteligência artificial não precisa necessariamente se parecer muito com a mente humana. IAs poderiam ser — e provavelmente a maioria delas será — completamente diferentes da inteligência humana. Devemos esperar que elas tenham arquiteturas cognitivas bastante distintas das inteligências biológicas, e em seus estágios iniciais de desenvolvimento elas terão perfis de fraquezas e poderes cognitivos bastante distintos (embora, como argumentaremos mais adiante, elas poderiam acabar superando qualquer deficiência inicial). Além do mais, os sistemas que definem os objetivos das IAs podem divergir completamente da maneira como a mente humana define objetivos. Não existe razão para esperarmos que uma IA genérica seja motivada por amor, ódio, orgulho ou qualquer outro sentimento comum aos seres humanos: essas adaptações complexas necessitariam de esforços custosos e deliberados para serem recriadas em uma IA. Esse é, ao mesmo tempo, um grande problema e uma grande oportunidade. Retornaremos à questão das motivações de uma IA em capítulos subsequentes, mas ela é tão central para o nosso argumento que vale a pena mantê-la em mente ao longo de todo o livro.

Emulação completa do cérebro

Em uma emulação completa do cérebro (também conhecida como *uploading*), softwares inteligentes seriam produzidos através do mapeamento e modelagem precisa da estrutura computacional de um cérebro biológico. Essa abordagem representa um caso extremo de inspiração na natureza: plágio descarado. Para se chegar a uma emulação completa do cérebro, é preciso seguir os seguintes passos.

Primeiro, um mapeamento suficientemente detalhado de um cérebro humano específico é criado. Isso pode envolver a estabilização *post mortem* do cérebro por vitrificação (processo que transforma o tecido em uma espécie de vidro). Uma máquina poderia então dissecar o tecido em fatias finas que seriam, por sua vez, colocadas em um escâner, o qual provavelmente obteria imagens por meio de uma série de microscópios eletrônicos. Durante esse estágio, várias substâncias de

contraste poderiam ser aplicadas de modo a ressaltar diferentes propriedades estruturais e químicas. Diversos escâneres poderiam trabalhar em paralelo a fim de processar diversas fatias de cérebro simultaneamente.

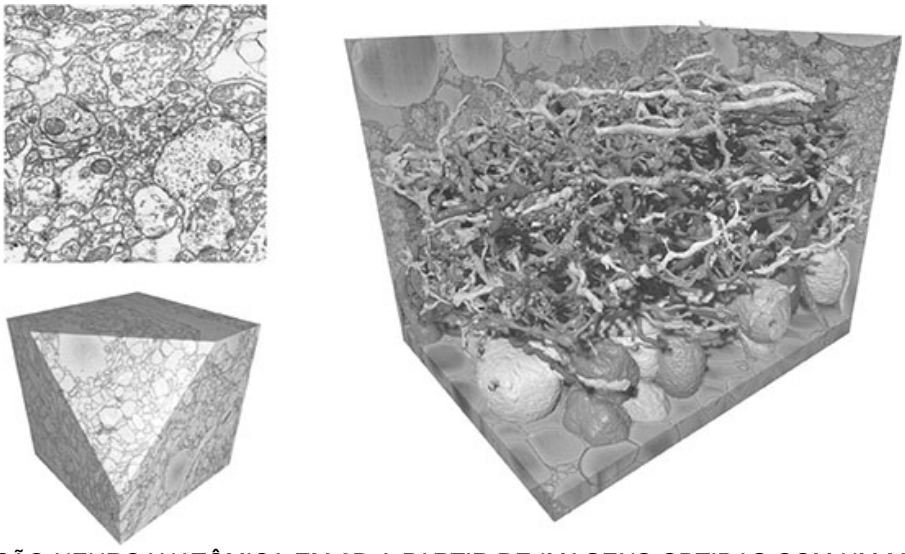
Segundo, os dados brutos obtidos a partir desses escâneres são inseridos em um computador para processamento automático das imagens com o objetivo de reconstruir o modelo tridimensional da rede neural que implementaria a cognição no cérebro original. Na prática, essa etapa poderia acontecer de forma concomitante com a primeira etapa para reduzir a quantidade de imagens de alta resolução armazenadas em *buffers*. O mapa resultante desse processo seria, então, combinado com uma biblioteca que contasse com diversos modelos neurocomputacionais de diferentes tipos de neurônios ou de diferentes elementos neurais (como alguns tipos específicos de conectores sinápticos). A figura 4 mostra alguns resultados produzidos com a tecnologia que possuímos atualmente para escaneamento e processamento de imagens.

No terceiro estágio, a estrutura neurocomputacional resultante dos dois estágios anteriores seria implementada em um computador suficientemente potente. Se bem-sucedida, o resultado seria uma reprodução digital do intelecto original com memória e personalidade intactas. A emulação de uma mente humana passaria então a existir como um software em um computador. Essa mente poderia tanto habitar uma realidade virtual como interagir com o mundo exterior através de apêndices robóticos.

A via da emulação completa do cérebro não requer a compreensão do funcionamento da cognição humana ou saber programar uma inteligência artificial. Ela requer apenas que tenhamos entendimento a respeito das características funcionais de baixo nível dos elementos computacionais do cérebro. Nenhuma descoberta fundamental teórica ou conceitual é necessária para que a emulação completa do cérebro seja bem-sucedida.

A emulação completa do cérebro necessita, entretanto, de algumas tecnologias avançadas para se tornar realidade. Existem três pré-requisitos principais: (1) *escaneamento*: microscopia de alto rendimento, com resolução suficiente e capacidade de detecção de propriedades relevantes; (2) *tradução*: análise automática de imagens que transforme dados brutos escaneados em um modelo interpretado tridimensional de elementos neurocomputacionais relevantes; e (3) *simulação*: hardware suficientemente potente para implementar a estrutura computacional resultante (ver tabela 4). (Em comparação com essas etapas mais desafiadoras, a construção de uma realidade virtual básica ou de um corpo robótico com entrada audiovisual e alguns canais de saída simples é relativamente fácil. Entrada e saída de dados simples, mas minimamente adequados, já se mostram factíveis com a tecnologia atual).[103]

FIGURA 4



RECONSTRUÇÃO NEUROANATÔMICA EM 3D A PARTIR DE IMAGENS OBTIDAS COM UM MICROSCÓPIO ELETRÔNICO. *À esquerda, acima:* típico micrográfico eletrônico mostrando um corte transversal de material neural — dendritos e axônios. *À esquerda, abaixo:* imagem volumétrica do tecido da retina neural de um coelho, feita com microscopia eletrônica de varredura de bloco-face de série.* Imagens 2D individuais foram sobrepostas de modo a formar um cubo (com o tamanho aproximado de 11 µm). *À direita:* reconstrução de um subconjunto de projeções neurais preenchendo o volume de um neurópilo, gerado por um algoritmo de segmentação automática.**
Helmstaedter et al. (2011). Andres et. al. (2012).*

Existe uma boa razão para imaginar que essas tecnologias necessárias como pré-requisito são possíveis de alcançar, embora não em um futuro próximo. Modelos computacionais razoáveis de diversos tipos de neurônios e processos neurais já existem. Também já foram desenvolvidos softwares de reconhecimento de imagem capazes de rastrear axônios e dendritos em meio a diversas imagens bidimensionais sobrepostas (embora sua confiabilidade precise ser aprimorada). Existem ainda ferramentas de captação de imagens que já fornecem a resolução necessária — usando um microscópio de tunelamento com varredura, já é possível "ver" átomos individuais, o que é uma resolução muito maior do que seria necessário. Entretanto, apesar das nossas capacidades e conhecimentos atuais sugerirem que não há, *a priori*, uma barreira para o desenvolvimento das tecnologias necessárias, fica claro que ainda precisaríamos de um grande progresso técnico incremental para atingir o ponto de efetuar uma emulação completa do cérebro.[104] Por exemplo, será necessário que a tecnologia de microscopia tenha não apenas resolução suficiente, mas também rendimento suficiente. Utilizar um microscópio de tunelamento com varredura de resolução atômica para obter as imagens necessárias da superfície seria um processo demasiadamente lento, o que o torna impraticável. Seria mais plausível a utilização de um microscópio eletrônico de menor resolução, mas, para que isso ocorra, seriam necessários novos métodos de preparação e contraste do tecido cortical para tornar visíveis detalhes relevantes, como a estrutura sináptica detalhada. Uma grande expansão das bibliotecas neurocomputacionais e melhorias significativas em processamento e interpretação automática de imagens também seriam necessárias.

TABELA 4

Capacidades necessárias para a emulação completa do cérebro

Escaneamento	Pré processamento/ fixação		Preparação apropriada do cérebro e retenção da microestrutura relevante e do seu estado
	Processamento físico		Métodos de manipulação do cérebro e tecidos fixos, antes, durante e depois do escaneamento
	Captação de imagens	Volume	Capacidade de escanear todo o volume do cérebro em um período de tempo e a um custo razoáveis
		Resolução	Escanear as imagens em uma resolução que seja suficiente para a sua reconstrução
		Informações funcionais	Capacidade do escaneamento de detectar propriedades funcionalmente relevantes de tecidos
Tradução	Processamento de imagens	Ajuste geométrico	Lidar com distorções criadas por imperfeições nas imagens escaneadas
		Interpolação de dados	Lidar com dados ausentes
		Redução de ruído	Aprimorar a qualidade das imagens escaneadas
		Rastreamento	Detectar estruturas e processá-las em um modelo tridimensional consistente do tecido
	Interpretação das imagens escaneadas	Identificação do tipo de célula	Identificar tipos de célula
		Identificação de sinapses	Identificar as sinapses e sua conectividade
		Estimativa de parâmetros	Estimar os parâmetros funcionalmente relevantes

			de células, sinapses e outras entidades
		Criação de base de dados	Armazenar esse conjunto de dados de forma eficiente
	Modelo de software do sistema neural	Modelo matemático	Modelagem de entidades e seus comportamentos
		Implementação eficiente	Implementação do modelo
Simulação	Armazenamento		Armazenamento do modelo original e seu estado atual
	Largura de banda		Comunicação eficiente entre processadores
	CPU		Poder de processamento para executar a simulação
	Simulação corporal		Simulação de um corpo que permita a interação com um ambiente virtual ou com um ambiente real
	Simulação do ambiente		Ambiente virtual para um corpo virtual

Em geral, a emulação completa do cérebro depende menos de um insight teórico e mais da capacidade tecnológica do que a inteligência artificial. Exatamente quanta tecnologia seria necessária para a emulação completa do cérebro depende do grau de abstração no qual o cérebro é emulado. Nesse sentido, há uma relação de troca entre insight e tecnologia. Em geral, quanto mais inferiores os equipamentos de aquisição de imagens e mais rudimentares os computadores, menos poderemos contar com a possibilidade de simular processos cerebrais químicos e eletrofisiológicos de baixo nível, e maior deverá ser nossa compreensão teórica da arquitetura computacional que estamos tentando emular, a fim de criar representações mais abstratas das funcionalidades relevantes.[105] Por outro lado, com uma tecnologia de aquisição de imagem suficientemente avançada e poder computacional abundante, talvez seja possível fazer uma emulação na base da força bruta, mesmo com uma compreensão limitada a respeito do cérebro. Em um caso limite irreal, é possível imaginar a emulação de um cérebro no nível de suas partículas elementares por meio da aplicação da equação de mecânica quântica de Schrödinger. Seria possível, então, basear-se inteiramente em conhecimentos de

física preexistentes e não usar nenhum modelo biológico. Esse caso extremo exigiria, entretanto, um poder computacional e uma aquisição de dados quase impraticável. Um nível muito mais plausível de emulação seria aquele que incorporasse neurônios individuais e suas matrizes de conectividade, além de parte da estrutura de suas árvores dendríticas e talvez algumas variáveis de estado de sinapses individuais. Moléculas neurotransmissoras não seriam simuladas individualmente, mas suas concentrações flutuantes seriam modeladas de forma aproximada.

Para avaliar a viabilidade da emulação completa do cérebro, é preciso compreender os critérios de sucesso. A intenção não é criar uma simulação do cérebro tão detalhada e precisa que seja possível prever exatamente o que aconteceria no cérebro original caso ele fosse submetido a uma determinada sequência de estímulos. Ao contrário, o objetivo é capturar uma quantidade suficiente de propriedades computacionalmente funcionais do cérebro para possibilitar que a emulação resultante seja capaz de realizar trabalho intelectual. Para tal propósito, muitos dos detalhes biológicos complicados de um cérebro real são irrelevantes.

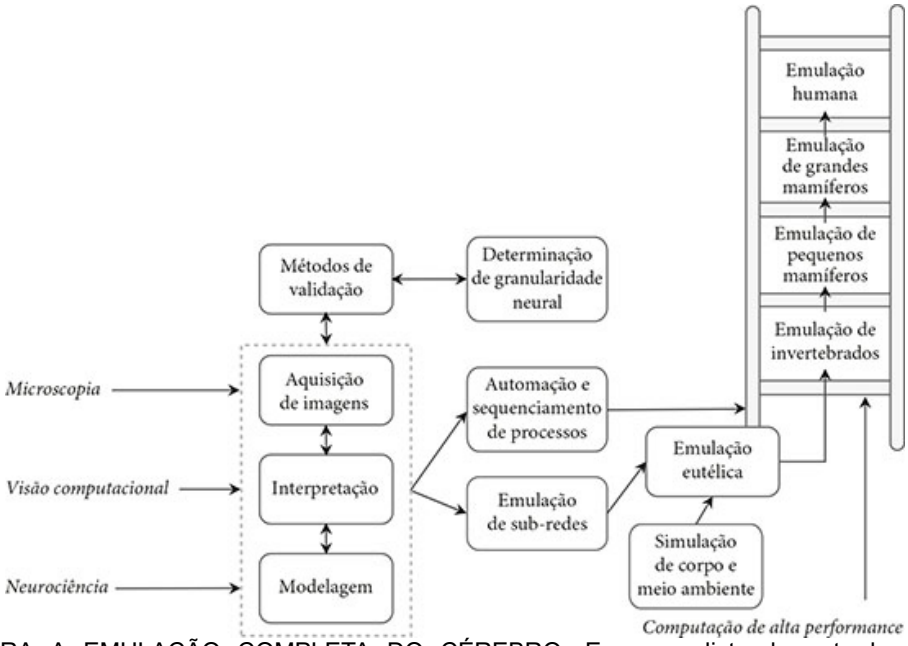
Uma análise mais elaborada poderia distinguir os diferentes níveis de sucesso na emulação, tendo como base o quão preservadas se mantiveram as funcionalidades de processamento de informação no cérebro emulado. Por exemplo, seria possível distinguir entre (1) uma *emulação de alta fidelidade*, que capture todos os aspectos referentes a conhecimento, habilidades, capacidades e valores do cérebro emulado; (2) uma e*mulação distorcida* cujas predisposições são significativamente não humanas em alguns aspectos, mas que tenha capacidade de efetuar praticamente o mesmo trabalho intelectual que o cérebro emulado; e (3) uma e*mulação genérica* (que também pode ser distorcida), que seria muito semelhante ao cérebro de um bebê, não capturando as memórias e habilidades adquiridas pelo cérebro adulto emulado, mas com a capacidade de aprender quase tudo que um humano normal também é capaz de aprender.[106]

Embora, em última instância, seja aparentemente possível fazer uma *emulação de alta fidelidade*, tudo indica que, caso sigamos por essa linha, a primeira emulação completa que seríamos capazes de fazer resultaria *distorcida*. Provavelmente, antes que sejamos capazes de fazer com que as coisas funcionassem perfeitamente, faríamos com que elas funcionassem de maneira imperfeita. É possível também que um esforço em direção às tecnologias de emulação nos levasse à criação de algum tipo de IA neuromórfica capaz de adaptar alguns dos princípios neurocomputacionais que fossem descobertos durante o processo de emulação, hibridizando-os então com métodos sintéticos, e isso aconteceria antes da finalização de uma emulação completa do cérebro totalmente funcional. Como veremos mais adiante, a possibilidade de tal processo desencadear em uma IA neuromórfica complica uma avaliação estratégica das eventuais vantagens de acelerar o desenvolvimento de tecnologias de emulação.

Atualmente, o quão longe estamos de alcançar uma emulação completa do cérebro humano? Uma avaliação recente apresentou um roteiro técnico e concluiu que as competências necessárias para tal feito podem estar disponíveis em meados

do século, embora tal previsão tenha uma considerável margem de erro.[107] A figura 5 apresenta as principais etapas desse roteiro. Não devemos nos enganar com a sua aparente simplicidade, mas ser cuidadosos em não subestimar o volume de trabalho que ainda deverá ser feito. Nenhum cérebro foi emulado ainda. Considere o humilde organismo da espécie *Caenorhabditis elegans*, um nematelminto transparente de aproximadamente 1 milímetro de comprimento e que possui 302 neurônios. A matriz completa de conectividade desses neurônios é conhecida desde meados da década de 1980, quando foi trabalhosamente mapeada através de dissecação, microscopia eletrônica e rotulagem manual dos espécimes.[108] Porém, simplesmente saber quais neurônios estão ligados entre si não é o suficiente. Para criar uma emulação cerebral, também seria necessário saber quais sinapses são estimuladoras e quais são inibidoras, compreender a força das conexões e conhecer as diversas propriedades dinâmicas dos axônios, sinapses e árvores de dendritos. Essas informações ainda não estão disponíveis sequer para o diminuto sistema nervoso dos *C. elegans* (embora atualmente é provável que isso seja passível de ser atingido através de um projeto de pesquisa específico de proporções razoáveis).[109] Ter sucesso na emulação de um cérebro minúsculo como o do *C. elegans* nos forneceria uma visão mais nítida a respeito do que seria necessário para emular cérebros maiores.

FIGURA 5



ROTEIRO PARA A EMULAÇÃO COMPLETA DO CÉREBRO. Esquema listando entradas, atividades e principais etapas.*

** Sandberg e Bostrom (2008). O relatório original apresenta explicações mais detalhadas.*

Em algum ponto durante o processo de desenvolvimento dessa tecnologia, a partir do momento em que técnicas automatizadas para a emulação de pequenas quantidades de tecido cerebral estiverem disponíveis, o problema passará a ser o dimensionamento. Perceba “a escada” no lado direito da figura 5. Essa série

ascendente de quadros representa uma sequência final de avanços que podem ser iniciados uma vez que sejam ultrapassados os obstáculos preliminares. Os estágios nessa sequência correspondem, sucessivamente, às emulações completas do cérebro de tipos de organismos cada vez mais sofisticados neurologicamente — por exemplo: *C. elegans* → *abelha* → *rato* → *macaco* → *ser humano*. Como as lacunas entre esses degraus — pelo menos a partir do primeiro — são, em sua maioria, quantitativas por natureza e geradas principalmente (mas não unicamente) pelas diferenças de tamanho dos cérebros que serão emulados, elas poderiam ser tratáveis com um aumento relativamente direto da capacidade técnica de escaneamento e de simulação.[110]

Uma vez que se tenha começado a subir essa "escada" final, a concretização última da emulação completa de um cérebro humano se tornaria algo muito mais previsível de acontecer num futuro próximo.[111] Seguindo o caminho da emulação completa do cérebro, devemos esperar algum tipo de aviso antes do surgimento de uma inteligência de máquina de nível humano, pelo menos se as últimas entre as tecnologias necessárias para alcançar o amadurecimento suficiente forem o escaneamento de alto desempenho ou o poder computacional necessário para realizar simulações em tempo real. Se, no entanto, a última tecnologia necessária for a modelagem neurocomputacional, então a transição de protótipos inexpressivos para uma emulação humana funcional poderia ser mais abrupta. É possível imaginar um cenário no qual, apesar da presença abundante de dados escaneados e de computadores rápidos, seja ainda difícil fazer com que os modelos neurais funcionem devidamente. Quando finalmente o último obstáculo for superado, o que era anteriormente um sistema completamente disfuncional — talvez semelhante a um cérebro inconsciente sofrendo uma convulsão — poderá alcançar abruptamente um estado alerta coerente. Nesse caso, o principal avanço não seria alcançado através de uma série de emulações funcionais de animais em ordem crescente de tamanho (que provocariam manchetes nos jornais a cada etapa atingida, com tamanhos de fonte cada vez maiores). Até mesmo para aqueles mais atentos, seria difícil prever quantas falhas estariam presentes nos modelos neurocomputacionais durante qualquer etapa ou quanto tempo seria necessário para corrigir essas falhas, mesmo às vésperas de um avanço significativo. (Uma vez que uma emulação completa de um cérebro humano fosse alcançada, novos desenvolvimentos potencialmente explosivos ocorreriam, mas iremos nos abster dessa discussão até o capítulo 4.)

É possível, então, imaginar cenários com muitas surpresas em relação à emulação completa do cérebro, ainda que todas as pesquisas relevantes fossem conduzidas abertamente. Contudo, em comparação ao caminho da IA, a emulação completa do cérebro é mais provável de ser precedida por sinais claros, uma vez que se ampara mais em tecnologias concretas e observáveis, e não totalmente em insights teóricos. Podemos também afirmar com muito mais confiança do que em relação à opção pela via da IA que a emulação não deverá acontecer em um futuro próximo (digamos nos próximos quinze anos), pois sabemos que diversas tecnologias precursoras ainda não foram desenvolvidas. Comparativamente, é plenamente possível que alguém seja capaz, *em princípio*, de desenvolver o código

de uma IA embrionária usando qualquer computador pessoal comum atual; e também é concebível, apesar de pouco provável, que alguém, em algum lugar, terá um insight de como fazer isso em um futuro muito próximo.

Cognição biológica

Um terceiro caminho para a criação de uma inteligência superior à humana é o aprimoramento da performance dos cérebros biológicos. Em princípio, isso poderia ser obtido através da reprodução seletiva, sem ajuda da tecnologia. Qualquer tentativa de iniciar um programa eugenista clássico em larga escala enfrentaria, entretanto, uma série de restrições políticas e morais. Além disso, a não ser que houvesse uma seleção muito intensa, seriam necessárias muitas gerações para que fosse possível produzir resultados substanciais. Muito antes que tal iniciativa gerasse resultados, avanços em biotecnologia permitiriam um controle muito mais direto da genética humana e da neurobiologia, tornando obsoleto qualquer programa de reprodução seletiva. Assim, focaremos aqui nos métodos que têm potencial de apresentar resultados mais rapidamente — no período de algumas poucas gerações ou menos.

Nossa capacidade cognitiva individual pode ser aperfeiçoada de diversas maneiras, incluindo métodos tradicionais, como educação e treinamento. O desenvolvimento neurológico pode ser promovido através de intervenções de baixo nível tecnológico, como melhorias na nutrição materna e infantil, remoção de chumbo e outros poluentes neurotóxicos do meio ambiente, erradicação de parasitas, incentivo a níveis adequados de sono e exercício, além da prevenção de doenças que afetam o cérebro.[112] Melhorias cognitivas podem, certamente, ser alcançadas por tais vias, mas os ganhos seriam provavelmente modestos, em especial nas populações já razoavelmente bem nutridas e com bons níveis educacionais. É certo que não alcançaremos a superinteligência por nenhum desses meios, mas eles podem auxiliar de maneira indireta, especialmente ao propiciar melhorias nas condições de populações despossuídas, o que propiciaria uma melhor captação global de talentos. (Uma diminuição constante nos níveis de inteligência ao longo da vida continua acontecendo em muitas regiões empobrecidas do planeta em virtude da deficiência de iodo — algo vergonhoso, uma vez que isso pode ser evitado por meio da fortificação do sal com iodo a um gasto anual de apenas alguns centavos por pessoa).[113]

A melhoria da biomedicina poderia permitir impulsos maiores. Atualmente já existem drogas que supostamente melhoram a memória, a concentração e a energia mental em pelo menos alguns aspectos.[114] (Este livro foi escrito na base do consumo de café e chicletes de nicotina.) Apesar da eficácia da atual geração de "drogas inteligentes" ser variável, marginal e, em geral, dúbia, futuramente os nootrópicos poderiam oferecer benefícios mais claros e um número menor de efeitos colaterais.[115] No entanto, parece pouco plausível, tanto do ponto de vista neurológico quanto do evolucionário, a possibilidade de aumentar drasticamente a inteligência de uma pessoa saudável apenas a partir da introdução de algum

elemento químico no cérebro desse indivíduo.[116] O funcionamento cognitivo do cérebro humano depende de um delicado equilíbrio entre diversos fatores, especialmente durante os estágios críticos do desenvolvimento embrionário — e é muito mais provável que, para ser melhorada, essa estrutura auto-organizável precise ser cuidadosamente balanceada, sintonizada e cultivada em vez de simplesmente receber a introdução de uma poção química estranha.

A manipulação da genética fornecerá ferramentas muito mais poderosas do que a psicofarmacologia. Considere mais uma vez a ideia de seleção genética: em vez de tentar implementar programas eugenistas através do controle de progenitores, poderíamos fazer uma seleção a nível embrionário ou de gametas.[117] Diagnósticos genéticos pré-implante já têm sido aplicados em procedimentos de fertilização *in vitro* (FIV) para rastrear, nos embriões produzidos, desordens monogênicas como a doença de Huntington e predisposição a outras doenças de início tardio, como o câncer de mama. Também têm sido usados para a escolha do sexo e para verificar a compatibilidade genética de leucócitos com os de um irmão doente, que pode, dessa forma, beneficiar-se de uma doação de células-tronco do cordão umbilical quando do nascimento do bebê.[118] A quantidade de características que poderão ser selecionadas para ser mantidas ou descartadas irá se expandir imensamente nas próximas duas décadas. Um fator que favorecerá o progresso da genética comportamental será a queda rápida dos custos de genotipagem e sequenciamento genético. A análise de características complexas envolvendo todo o genoma a partir de estudos que se utilizam de um grande número de indivíduos é, atualmente, algo cada vez mais factível, o que aumentará em muito nosso conhecimento a respeito da arquitetura genética da cognição e comportamento humanos.[119] Qualquer característica que possa ser herdada de alguma forma — incluindo as capacidades cognitivas — poderia, então, ser suscetível à seleção.[120] A seleção de embriões não requer um entendimento profundo acerca dos caminhos causais pelos quais os genes em suas complexas relações com o meio ambiente produzem fenótipos: requer apenas dados (muito numerosos) a respeito da correlação genética das características desejadas.

É possível fazer uma estimativa do grau de magnitude dos ganhos possíveis que podem ser obtidos com a seleção em diferentes cenários.[121] A tabela 5 mostra os aumentos em inteligência esperados resultantes de seleções em conjuntos de diversos tamanhos, caso se possua total conhecimento a respeito das variantes genéticas aditivas comuns subjacentes à herdabilidade da inteligência no sentido restrito. (Com informações parciais, a eficácia da seleção seria reduzida, embora não tanto quanto poderíamos, ingenuamente, supor.)[122] Não surpreende que fazer uma seleção entre números maiores de embriões produza ganhos maiores, mas devemos notar que a vantagem com o aumento diminui acentuadamente, ou seja, há uma diminuição do retorno: a seleção entre cem embriões não produz um ganho nem perto de cinquenta vezes o que seria obtido a partir de uma seleção realizada com dois embriões.[123]

TABELA 5

Ganhos máximos de QI a partir da seleção de embriões em conjuntos de diferentes tamanhos*

Seleção	Pontos de QI ganhos
1 em 2	4,2
1 em 10	11,5
1 em 100	18,8
1 em 1000	24,3
5 gerações de 1 em 10	< 65 (devido à diminuição de retorno)
10 gerações de 1 em 10	< 130 (devido à diminuição de retorno)
Limites cumulativos (variantes aditivas otimizadas para cognição)	100+ (< 300 (devido à diminuição de retorno))

**Essa tabela foi retirada de Shulman e Bostrom (2014). Ela é baseada em um modelo simplificado que assume uma distribuição gaussiana de QIs previstos para os embriões com um desvio padrão de 7,5 pontos. A quantidade de aumento cognitivo que seria alcançada com diferentes números de embriões depende do quão diferentes esses embriões são entre si em relação às variantes genéticas aditivas cujos efeitos são conhecidos. Irmãos têm um coeficiente de parentesco de ½, e variantes genéticas aditivas comuns são responsáveis por menos da metade da variação da inteligência fluida nos adultos (Davies et al., 2011). Esses dois fatores sugerem que onde o desvio padrão da população observado em países desenvolvidos é de 15 pontos, o desvio padrão de influências genéticas presentes num grupo de embriões será de 7,5 pontos ou menos.*

Curiosamente, essa diminuição de retorno é bastante reduzida quando a seleção é distribuída entre múltiplas gerações. Nesse caso, selecionar repetidamente o melhor entre dez embriões ao longo de dez gerações (onde cada nova geração consiste dos filhos daqueles já selecionados na geração anterior) resultará em um aumento muito maior do valor das características do que uma única seleção de 1 em 100. O problema com as seleções sequenciais é que elas demoram mais. Se cada degrau geracional precisa de vinte a trinta anos, então uma seleção com apenas cinco gerações sucessivas terminaria, tranquilamente, apenas no século XXII. Bem antes disso, meios muito mais diretos e poderosos de engenharia genética (isso sem mencionar a inteligência de máquina) provavelmente já estarão disponíveis.

Existe, entretanto, uma tecnologia complementar, a qual, uma vez desenvolvida para o uso em humanos, poderia potencializar o poder do rastreamento genético pré-implantação: a derivação de óvulos e espermatozoides viáveis a partir de células-tronco embrionárias.[124] As técnicas para a realização desse processo já foram utilizadas na produção de filhotes férteis em ratos e células semelhantes a gametas em seres humanos. Ainda restam, entretanto, alguns desafios científicos

substanciais, entre eles a questão de como traduzir os resultados em animais para a aplicação em seres humanos e de que maneira poderiam ser evitadas anormalidades epigenéticas nas linhas de células-tronco resultantes. De acordo com uma especialista, esses desafios podem fazer com que sua aplicação humana só aconteça daqui a "dez ou até mesmo cinquenta anos".[125]

Com gametas derivados de células-tronco, o tamanho do poder de seleção disponível para um casal seria imensamente aumentado. Com as práticas utilizadas atualmente, um procedimento de fertilização *in vitro* envolve tipicamente a criação de menos de dez embriões. Com gametas derivados de células-tronco, a doação de algumas poucas células poderia resultar em um número praticamente ilimitado de gametas que poderiam ser combinados para produzir embriões, que, por sua vez, poderiam ser submetidos à genotipagem ou sequenciados e, a partir disso, seria possível selecionar o mais promissor para o implante. Dependendo do custo de preparação e triagem de cada embrião, essa tecnologia poderia gerar um enorme aumento do poder de seleção disponível para casais com a utilização da fertilização *in vitro*.

Mas, ainda mais importante, as células-tronco derivadas de gametas permitiriam que múltiplas gerações de seleção fossem comprimidas em um período menor do que o da maturação humana, uma vez que permitiriam um *processo iterativo de seleção de embriões*. Esse procedimento consistiria das seguintes etapas:[126]

1. Tipificar geneticamente e selecionar um número de embriões com maior concentração de características genéticas desejadas.
2. Extrair células-tronco desses embriões e converter essas células em óvulos e espermatozoides, com amadurecimento de seis meses ou menos.[127]
3. Cruzar esses óvulos e espermatozoides para produzir embriões.
4. Repetir esse processo até que mudanças genéticas significativas tenham sido acumuladas.

Dessa forma, seria possível realizar a seleção referente a dez ou mais gerações em apenas alguns poucos anos. (Esse procedimento teria um custo alto de tempo e de dinheiro; no entanto, em princípio, ele precisaria ser feito apenas uma única vez, e não a cada nascimento. As linhas de células estabelecidas ao final do procedimento poderiam ser usadas para gerar um grande número de embriões melhorados.)

Conforme indicado na tabela 5, o nível *médio* de inteligência dos indivíduos concebidos a partir desse processo poderia ser muito alto, possivelmente igual ou até mesmo superior ao do indivíduo mais inteligente da história da humanidade. Um mundo que tivesse uma grande população formada por tais indivíduos poderia (se tivesse cultura, educação, infraestrutura de comunicação etc. em conformidade com o nível da população) constituir uma coletividade superinteligente.

O impacto dessa tecnologia será atenuado e retardado por diversos fatores. Existe o inevitável atraso no amadurecimento até que os embriões selecionados se tornem humanos adultos: levaria, pelo menos, vinte anos para que essas crianças aperfeiçoadas se tornassem produtivas e ainda mais tempo até que elas se tornassem um segmento substancial da força de trabalho. Além disso, mesmo após

o aperfeiçoamento dessa tecnologia, as taxas de adoção seriam provavelmente baixas a princípio. Alguns países poderiam, inclusive, proibir, com base em princípios morais ou religiosos, o uso dessa tecnologia.[128] Mesmo em lugares onde a seleção de embriões fosse permitida, muitos casais poderiam preferir a concepção natural. Contudo, a disposição para o uso da FIV aumentaria na medida em que fossem observados os evidentes benefícios associados ao procedimento — como a garantia virtual de que a criança seria altamente talentosa e sem predisposição genética a doenças. Um gasto menor com saúde e uma expectativa de remunerações maiores ao longo da vida seriam alguns dos benefícios que contariam a favor de uma escolha pela seleção genética. Conforme o uso do procedimento fosse se tornando mais comum, principalmente entre as elites, poderia surgir uma mudança cultural para normas parentais que considerassem o uso da seleção genética como algo possível para pais responsáveis e esclarecidos. Muitos daqueles que tivessem sido inicialmente relutantes poderiam aderir ao procedimento para que seus filhos não tivessem desvantagens diante das crianças aperfeiçoadas, filhas de seus colegas e amigos. Possivelmente, alguns países ofereceriam incentivos que encorajassem seus cidadãos a optar pelas vantagens da seleção genética com o objetivo de aumentar o capital de sua população, ou mesmo para aprimorar, a longo prazo, a estabilidade social por meio da seleção de características genéticas como docilidade, obediência, submissão, conformismo, aversão ao risco ou covardia em indivíduos que não pertencessem às classes dominantes.

Os efeitos na capacidade intelectual também dependeriam da extensão na qual o poder de seleção disponível fosse usado para aprimorar características cognitivas (tabela 6). Aqueles que optassem por utilizar alguma forma de seleção embrionária teriam que escolher a maneira de alocar o poder de seleção que estaria à disposição, e a inteligência estaria até certo ponto em competição com outros atributos igualmente desejáveis, como saúde, beleza, personalidade ou capacidade atlética. Ao oferecer uma enorme capacidade seletiva, o processo iterativo de seleção de embriões diminuiria um pouco a necessidade de escolha, uma vez que tornaria viável uma intensa seleção simultânea por múltiplas características. No entanto, esse procedimento teria uma tendência de romper a relação genética natural entre pais e filhos, o que poderia afetar negativamente sua demanda em diversas culturas.[129]

TABELA 6

Possíveis impactos da seleção genética em diferentes cenários*

Adoção / tecnologia	**“FIV+” Seleção de 1 em 2 embriões [4 pontos]**	**“FIV agressiva” Seleção de 1 em 10 embriões [12 pontos]**	**“Ovo in vitro” Seleção de 1 em 100 embriões [19 pontos]**	**“Processo iterativo de seleção de embriões” [100 pontos ou mais]**
“Prática de fertilização	Socialmente negligenciável ao	Socialmente negligenciável ao	Indivíduos gerados com	Indivíduos gerados com o processo de

incipiente” ~0,25% de adoção	longo de uma geração. Efeitos sociais controversos mais importantes do que os impactos diretos.	longo de uma geração. Efeitos sociais controversos mais importantes do que os impactos diretos.	o processo de seleção formam uma minoria notável em áreas que exigem alto desempenho intelectual.	seleção têm presença dominante entre os mais proeminentes cientistas, advogados, médicos, engenheiros. Renascença Intelectual?
“Vantagem para a elite” 10% de adoção	Pequeno impacto cognitivo na 1ª geração, combinado com uma seleção de características não cognitivas para aprimorar perceptivelmente uma minoria.	Uma grande parcela dos estudantes de Harvard aparenta melhorias. A 2ª geração é maioria nas profissões que demandam alta cognição.	Indivíduos gerados pelo processo de seleção tem presença dominante entre os cientistas, advogados, médicos e engenheiros, já na 1ª geração.	“Pós-humanidade”**
“Novo normal” > 90% de adoção	Distúrbios de aprendizado muito menos frequentes em crianças. Na 2ª geração, a população com QI alto quase dobrou.	Considerável aumento no nível de escolaridade e renda. A 2ª geração apresenta um substancial aumento entre os menos privilegiados.	QIs típicos de cientistas proeminentes passam a ser dez vezes mais comuns na 1ª geração. E milhares de vezes mais comuns na 2ª geração.	“Pós-humanidade”

*Tabela adaptada de Shulman e Bostrom (2014). **Bostrom (2008b).

Com avanços futuros em tecnologia genética, talvez se torne possível a sintetização de genomas de acordo com uma especificação, evitando a necessidade de um amplo banco de embriões. A síntese de DNA já é um processo de biotecnologia rotineiro e, em grande parte, automatizado, apesar de ainda não ser factível a sintetização completa de um genoma humano, de maneira que possa ser utilizado em um contexto reprodutivo (principalmente em virtude do fato de que ainda existem problemas não solucionados relacionados à epigenética).[130] Mas uma vez que essa tecnologia tenha alcançado um estágio mais avançado, seria possível desenvolver um embrião com a exata combinação desejada de material genético de ambos os pais. Poderiam ser introduzidos, também, genes que não estão presentes em nenhum dos pais, incluindo alelos de baixa frequência na população, mas que poderiam vir a ter um significativo impacto positivo na cognição.[131]

Quando for possível sintetizar o genoma humano, outra intervenção possível será a “correção” genética de um embrião. (O processo iterativo de seleção de embriões também poderia permitir algo semelhante.) Atualmente, cada um de nós possui uma carga mutacional composta, possivelmente, de centenas de mutações que reduzem a eficiência de inúmeros processos celulares.[132] Individualmente, cada mutação tem um efeito praticamente insignificante (e por isso é retirada lentamente da carga genética), mas ainda assim, se combinadas, tais mutações podem ter um pesado

impacto negativo em nosso funcionamento.[133] Diferenças individuais na inteligência podem, em grande medida, ser atribuídas a variações na natureza e no número de tais alelos levemente prejudiciais que cada um de nós carrega. Com a sintetização genética, poderíamos construir, a partir do genoma de um embrião, uma versão genética livre das distorções genéticas acumuladas pelas múltiplas mutações. Sendo um pouco mais provocativo, poderíamos dizer que indivíduos criados a partir desse genoma "corrigido" poderiam vir a ser "mais humanos" do que qualquer pessoa viva hoje em dia, uma vez que eles seriam expressões menos distorcidas da forma humana. Tais pessoas não seriam todas cópias idênticas umas das outras, uma vez que os seres humanos possuem variações entre si que vão muito além das mutações prejudiciais. Mas a manifestação no fenótipo de um genoma "corrigido" poderia ser uma constituição física e mental excepcional, com um funcionamento elevado de características poligênicas, como inteligência, saúde, determinação e aparência.[134] (É possível fazer uma analogia com rostos produzidos a partir da composição de fotografias individuais, onde os defeitos dos indivíduos sobrepostos são eliminados: ver figura 6.)

Outras técnicas potenciais da biotecnologia também poderiam ser relevantes. A clonagem humana, no momento em que fosse viável, poderia ser usada para replicar o genoma de indivíduos excepcionalmente talentosos. A aceitação de tal técnica seria limitada pela preferência dos futuros pais de possuírem uma ligação biológica com seus filhos, todavia, a prática da clonagem poderia, mesmo assim, ter um impacto relevante pelos seguintes motivos: (1) até mesmo um pequeno incremento no número de indivíduos excepcionalmente talentosos poderia apresentar um efeito significativo e (2) é possível que alguma nação adote um programa eugênico em larga escala, talvez a partir da contratação de mães de aluguel. Outros tipos de engenharia genética também podem se tornar importantes ao longo do tempo — como o desenvolvimento de genes sintéticos ou a inserção no genoma de regiões promotoras e outros elementos capazes de controlar a expressão gênica. Possibilidades ainda mais exóticas podem vir a existir, como recipientes repletos de tecido cortical complexo artificialmente cultivado ou animais transgênicos "aperfeiçoados" (talvez algum mamífero com cérebro grande, como a baleia ou o elefante, enriquecido com genes humanos). Esses últimos exemplos são puramente especulativos, mas, pensando a longo prazo, talvez não devêssemos descartá-los por completo.

Até aqui, discutimos intervenções na linha germinal, ou seja, aquelas que poderiam ser feitas a partir de gametas ou embriões. Aperfeiçoamentos genéticos somáticos gerados sem a necessidade de aguardar a finalização de ciclos geracionais poderiam, em princípio, produzir impactos mais rapidamente. Entretanto, eles são tecnologicamente muito mais desafiadores. Tais processos exigem que os genes modificados sejam inseridos em um grande número de células de um organismo vivo — incluindo o cérebro, no caso de aperfeiçoamentos cognitivos. A seleção de óvulos ou embriões já existentes, ao contrário, não requer a inserção de nenhum gene. Até mesmo tais terapias em linhas germinais que envolvem a modificação do genoma (como corrigir o genoma ou unir alelos raros) são muito mais fáceis de implementar nos estágios embrionários ou de gametas,

nos quais lidamos com um número reduzido de células. Ademais, intervenções na linha germinal de embriões provavelmente produziriam efeitos muito mais significativos do que intervenções somáticas em adultos, pois as primeiras seriam capazes de moldar o desenvolvimento do cérebro em seus estágios iniciais, enquanto as últimas poderiam apenas fazer alguns ajustes numa estrutura existente. (Algumas das mudanças possíveis através de terapias genéticas somáticas também poderiam ser alcançadas por meio da farmacologia.)

FIGURA 6



COMPOSIÇÃO DE ROSTOS COMO UMA METÁFORA PARA A CORREÇÃO DE GENOMAS. Cada uma das imagens centrais foi obtida a partir da sobreposição de fotografias de dezesseis indivíduos diferentes (moradores de Tel Aviv). Rostos obtidos a partir dessa sobreposição geralmente são considerados mais bonitos do que os rostos individuais que os originaram, já que as imperfeições idiossincráticas são eliminadas. De forma análoga, ao remover mutações individuais, os genomas corrigidos poderiam produzir pessoas mais próximas de um "ideal platônico". Tais indivíduos não seriam geneticamente idênticos, já que muitos genes existem em múltiplos alelos igualmente funcionais. A correção apenas eliminaria as variações oriundas de mutações prejudiciais.*

** Essas imagens compostas foram criadas por Mike Mike a partir de fotografias individuais tiradas por Virtual Flavius (Mike, 2013).*

Sendo assim, com o foco em intervenções na linha germinal, é necessário levar em consideração o intervalo geracional que atrasaria qualquer impacto substancial no mundo.[135] Mesmo que tal tecnologia já estivesse aperfeiçoada e fosse colocada em uso imediatamente, seriam necessárias mais de duas décadas para que os embriões geneticamente aperfeiçoados alcançassem a maturidade. Além disso, em procedimentos com seres humanos, há normalmente um intervalo de pelo menos uma década entre uma prova de conceito no laboratório e a aplicação clínica em virtude dos extensos estudos necessários para garantir a segurança do procedimento. As formas mais simples de seleção genética, porém, poderiam fazer com que tais testes fossem desnecessários, uma vez que a seleção genética poderia utilizar técnicas padrões de fertilização e informações genéticas para escolher entre embriões que seriam, de outra forma, selecionados aleatoriamente.

Atrasos também poderiam ser decorrentes de obstáculos oriundos não do medo do fracasso (exigência de testes de segurança), mas do medo do sucesso — necessidade de regulamentação em virtude de questões relacionadas à permissibilidade da seleção genética ou às suas implicações sociais mais amplas. Tais questões provavelmente influenciarão mais alguns países do que outros, considerando os diferentes contextos históricos, culturais e religiosos. A Alemanha

do pós-guerra, por exemplo, abriu mão de utilizar qualquer técnica reprodutiva que visasse, mesmo que remotamente, qualquer tipo de aperfeiçoamento genético, algo bastante compreensível diante da história particularmente sombria de atrocidades ligadas a movimentos eugenistas no país. É provável que outras nações ocidentais tenham uma postura mais liberal. E alguns países — talvez a China ou Cingapura, ambos com políticas populacionais de longo prazo — poderiam não apenas permitir, mas até mesmo estimular o uso da seleção genética e da engenharia genética para aprimorar a inteligência de suas populações quando essa tecnologia estiver disponível.

Uma vez que o exemplo inicial tenha sido dado e os resultados começarem a aparecer, aqueles que tiverem se mostrado inicialmente resistentes à ideia terão grandes incentivos para aderir a ela. Nações que ficassem de fora correriam um sério risco de se tornar defasadas cognitivamente e experimentariam enormes prejuízos econômicos, científicos e militares, perdendo prestígio frente a competidores que abraçassem as novas tecnologias de aperfeiçoamento humano. As pessoas passariam a ver as vagas em escolas de elite sendo preenchidas por crianças geneticamente selecionadas (que também poderiam ser mais bonitas, mais saudáveis e mais cuidadosas) e desejariam que seus filhos tivessem as mesmas vantagens. Uma vez que as tecnologias de aperfeiçoamento humano provassem ser efetivas e passassem a apresentar benefícios substanciais, existiria a possibilidade de que uma grande mudança de atitude viesse a ocorrer em um período de tempo relativamente curto, talvez algo em torno de uma década. Pesquisas de opinião nos Estados Unidos mostraram um aumento dramático da aprovação da FIV depois do nascimento do primeiro “bebê de proveta”, Louise Brown, em 1978. Alguns anos antes, apenas 18% dos norte-americanos disseram que utilizariam a FIV para tratar a infertilidade; mas em uma pesquisa feita logo após o nascimento de Louise Brown, 53% afirmaram que fariam o mesmo, e esse número só tem aumentado.[136] (Para efeito de comparação, em uma pesquisa feita em 2004, 28% dos norte-americanos aprovavam a seleção de embriões com o objetivo de “aprimorar a inteligência”, 58% a aprovavam como forma de evitar o aparecimento de câncer em adultos e 68% a aprovavam como meio de evitar doenças fatais em crianças.)[137]

Levando todos esses atrasos em consideração — digamos cinco a dez anos para que fosse possível juntar as informações necessárias para uma seleção significantemente efetiva a partir de um conjunto de embriões obtidos através de FIV (provavelmente muito antes que gametas provenientes de células-tronco estejam disponíveis para o uso na reprodução humana), dez anos para a captação de embriões e de vinte a 25 anos para que essa geração de indivíduos aperfeiçoados alcance a idade reprodutiva —, percebemos ser pouco provável que melhorias na linha germinal venham a ter algum impacto significativo na sociedade antes de meados deste século. Daí em diante, entretanto, a inteligência de segmentos significativos da população adulta pode começar a ser aprimorada geneticamente. A velocidade no aumento da inteligência seria intensamente acelerada conforme indivíduos concebidos através de novas tecnologias genéticas mais poderosas

(particularmente de gametas provenientes de células-tronco e do processo iterativo de seleção de embriões) entrassem no mercado de trabalho.

Uma vez que as tecnologias genéticas descritas acima estejam totalmente desenvolvidas (deixando de lado aqui as possibilidades mais exóticas, como inteligência desenvolvida em tecidos neurais cultivados artificialmente), talvez seja possível assegurar que esses novos indivíduos fossem, em média, mais inteligentes do que qualquer ser humano que já tenha existido, sendo que alguns deles atingiriam patamares de inteligência ainda maiores. Dessa forma, é possível afirmar que, em última instância, o potencial do melhoramento genético é alto, provavelmente o suficiente para que atingíssemos, pelo menos, formas mais fracas de superinteligência. Isso não deveria surpreender, afinal processos evolutivos aleatórios amplificaram dramaticamente a inteligência da linhagem humana, mesmo se comparada a dos nossos parentes mais próximos, os grandes primatas, ou a dos nossos ancestrais hominídeos; não existe razão para supor que o *Homo sapiens* atingiu o ápice da capacidade de cognição possível em um sistema biológico. Longe de sermos a espécie biologicamente possível mais inteligente, é melhor nos enxergarmos como, provavelmente, a mais estúpida das espécies biológicas capaz de iniciar uma civilização tecnológica — um nicho que preenchemos porque chegamos primeiro, e não por estarmos perfeitamente adaptados a ele.

O progresso pelo caminho biológico é algo claramente viável. O intervalo geracional presente em intervenções na linha germinal significa que o progresso não seria tão abrupto como numa conjuntura que envolvesse inteligência de máquina. (Terapias genéticas somáticas e intervenções farmacológicas poderiam, em teoria, evitar esse intervalo geracional, mas são tecnologias difíceis de ser aperfeiçoadas e é ainda menos provável que produzam algum efeito considerável.) O potencial *máximo* de uma máquina inteligente é, claramente, muito superior ao da inteligência orgânica. (É possível visualizar o quão grande é essa divergência se levarmos em consideração a diferença de velocidade entre componentes eletrônicos e células nervosas: atualmente, os transistores já são capazes de operar numa velocidade 10 milhões de vezes mais rápida que a dos neurônios biológicos.) Todavia, melhorias, mesmo que moderadas, na
capacidade biológica de cognição poderiam ter importantes consequências — como acelerar o desenvolvimento científico e tecnológico, incluindo os progressos em direção a formas mais aprimoradas de amplificação da inteligência biológica e da inteligência de máquina. Imagine qual seria a velocidade do progresso na área de inteligência artificial num mundo em que um joão-ninguém qualquer tivesse a mesma capacidade intelectual de Alan Turing ou John von Neumann e onde milhões de pessoas fossem mais inteligentes do que qualquer gênio do passado.[138]

Uma discussão acerca das implicações estratégicas do aperfeiçoamento cognitivo será tratada mais adiante, em outro capítulo. Mas é possível resumir essa seção salientando três conclusões: (1) por meio de melhorias biotecnológicas, é possível alcançar ao menos formas mais fracas de superinteligência; (2) a viabilidade de seres humanos cognitivamente aperfeiçoados torna mais plausível que formas avançadas de inteligência de máquina sejam viáveis — porque, mesmo

que para *nós* fosse fundamentalmente impossível criar uma inteligência de máquina (apesar de não existir nenhuma razão para supor isso), é possível que seres humanos aperfeiçoados consigam desenvolvê-la; e (3) quando considerados cenários que se estendem para além da segunda metade deste século, devemos levar em conta o provável surgimento de uma geração de indivíduos geneticamente aperfeiçoados — eleitores, inventores, cientistas —, com o grau de aperfeiçoamento aumentando rapidamente nas décadas subsequentes.

Interfaces cérebro-computador

Já foi sugerido algumas vezes que interfaces cérebro-computador diretas, principalmente implantes, poderiam tornar os seres humanos capazes de se aproveitar dos pontos fortes de um computador — como memória, rapidez e precisão na realização de cálculos aritméticos e altas taxas de transmissão de dados —, fazendo com que o sistema híbrido resultante superasse em muito um cérebro normal.[139] Porém, embora a possibilidade de uma conexão direta entre um cérebro humano e um computador já tenha sido demonstrada, é pouco provável que tais interfaces venham a ser largamente utilizadas como forma de aperfeiçoamento de seres humanos num futuro próximo.[140]

Para começo de conversa, existem significativos riscos de complicações médicas — infecções, deslocamento de eletrodos, hemorragia e declínio cognitivo — quando um eletrodo é implantado no cérebro. Até hoje, talvez o melhor exemplo dos benefícios que podem ser obtidos através do estímulo cerebral seja o tratamento de pacientes com mal de Parkinson. O implante para Parkinson é relativamente simples: não existe de fato uma comunicação com o cérebro, o aparelho apenas fornece uma corrente elétrica estimulante ao núcleo subtalâmico. Um vídeo de demonstração mostra um indivíduo caído em uma cadeira, completamente imóvel em decorrência da doença e, de repente, ele é trazido de “volta à vida” quando a corrente elétrica é acionada: o indivíduo passa a mover os braços, se levanta e anda até o outro lado do quarto, vira-se e dá uma pirueta. Entretanto, mesmo esse procedimento simples e quase milagroso tem seu lado negativo. Um estudo de pacientes com mal de Parkinson que receberam implantes cerebrais mostrou que, comparados àqueles que não possuem implantes, foram observadas reduções na fluência verbal, atenção seletiva, capacidade de nomear cores e memória verbal. Esses pacientes também apresentaram mais queixas em relação a problemas cognitivos.[141] Tais riscos e efeitos colaterais podem ser toleráveis se o procedimento é realizado com o intuito de aliviar os sintomas e a progressão de uma doença grave. Mas para que indivíduos saudáveis resolvessem ser candidatos a neurocirurgias, seria preciso que a melhoria da funcionalidade normal obtida a partir desse procedimento fosse bastante substancial.

Isso nos leva ao segundo motivo para duvidar de que a superinteligência será alcançada através da “ciborguização” — é muito provável que obter melhorias cognitivas seja algo muito mais complicado do que seu uso terapêutico. Pacientes que sofrem de paralisia podem se beneficiar de implantes que substituam os nervos danificados ou que ativem os padrões geradores de movimento da espinha

dorsal.[142] Pacientes cegos ou surdos podem se beneficiar de cócleas e retinas artificiais.[143] Pacientes com mal de Parkinson ou dores crônicas podem se beneficiar de estímulos cerebrais que inibam ou estimulem a atividade numa região particular do cérebro.[144] O que parece ser muito mais difícil é obter uma alta taxa de interação entre cérebro e computador que seja capaz de gerar um aumento significativo na inteligência que não pudesse ser mais facilmente alcançado por outros meios. A maioria dos benefícios que poderiam potencialmente ser obtidos com implantes cerebrais em indivíduos saudáveis poderia ser conseguida de forma muito menos arriscada, dispendiosa e inconveniente, por meio do uso de nossos órgãos motores e sensoriais na interação com computadores localizados fora de nossos corpos. Não é necessário conectar um cabo de fibra ótica em nosso cérebro para que tenhamos acesso à internet. A retina humana não só é capaz de transmitir dados numa impressionante velocidade de quase 10 milhões de bits por segundo como já vem equipada com uma enorme quantidade de *wetware* dedicado, o córtex visual, que é altamente adaptado para extrair significado desse grande fluxo de informações e fazer uma interface com outras áreas do cérebro que posteriormente processarão os dados.[145] Mesmo que exista uma maneira simples de injetar mais informação em nossos cérebros, esses dados extras teriam pouca influência na rapidez com que pensamos e aprendemos, a menos que todas as regiões neurais responsáveis pelo processamento dos dados sofressem um *upgrade* semelhante. Uma vez que isso incluiria praticamente todo o cérebro, seria necessária uma "prótese completa do cérebro" — o que nada mais é do que inteligência artificial geral. Caso tivéssemos IA de nível humano, neurocirurgias desse tipo não seriam necessárias: é indiferente se um computador possui um revestimento de metal ou de ossos. Sendo assim, esse caso limite nos leva de volta ao caminho da IA, o qual já examinamos.

Interfaces cérebro-computador também foram propostas como uma forma de extrair informações do cérebro para permitir a comunicação com outros cérebros ou com máquinas.[146] Tais ligações têm ajudado pacientes com síndrome do encarceramento a se comunicar com o mundo exterior ao permitir que eles movam um cursor por uma tela através do pensamento.[147] A largura de banda obtida em tais experimentos é baixa: o paciente consegue, a duras penas, digitar uma letra após a outra a uma velocidade de poucas palavras por minuto. É possível facilmente imaginar versões mais desenvolvidas dessa tecnologia — talvez a próxima geração de implantes possa se conectar com a área de Broca (uma região no lobo frontal que atua na produção de linguagem) para captar a fala interna.[148] Mas mesmo que essa tecnologia pudesse auxiliar pessoas com algum tipo de deficiência decorrente de um derrame ou degeneração muscular, ela teria pouco a oferecer a indivíduos saudáveis. Sua funcionalidade seria muito semelhante à de um microfone ligado a um software de reconhecimento de voz, algo que já está disponível comercialmente — e não vem acompanhado de dores, inconveniências, custos e riscos associados a uma neurocirurgia (além de não ter o caráter hiperorwelliano que seria associado a um dispositivo de escuta intracraniano). Manter as máquinas fora de nosso corpo também torna mais fácil fazer upgrades.

Mas e se não fosse sequer necessário usar palavras e fôssemos capazes de estabelecer uma conexão entre dois cérebros que permitisse transmitir conceitos, pensamentos ou fazer o "download" de toda uma área do conhecimento de uma mente para a outra? É possível fazer o download de extensos arquivos em nossos computadores, até mesmo de bibliotecas inteiras de livros e artigos, e isso pode ser feito em questão de segundos: será que algo similar não poderia ser feito com nossos cérebros? A aparente plausibilidade dessa ideia vem, muito provavelmente, de uma visão incorreta a respeito de como a informação é armazenada e representada em nosso cérebro. Conforme já foi dito aqui, a principal barreira que limita a inteligência humana não é a velocidade com a qual dados brutos são inseridos no cérebro, mas a velocidade com a qual o cérebro consegue extrair significado e sentido desses dados. Talvez venha a ser sugerido que a transmissão de sentido seja feita de forma direta, e não transformada em dados sensoriais que necessitem ser decodificados pelo recipiente. Existem dois problemas em relação a isso. O primeiro é que o cérebro, diferentemente dos programas que costumamos rodar em nossos computadores, não tem um padrão único de armazenamento e formato de dados. Ao contrário, cada cérebro desenvolve suas próprias representações idiossincráticas de conteúdos complexos. O grupo de neurônios que é recrutado para representar um conceito em particular depende exclusivamente das experiências de cada cérebro em questão (além de inúmeros fatores genéticos e processos estocásticos da nossa fisiologia). Assim como nas redes neurais artificiais, o sentido das coisas em uma rede neural biológica é, provavelmente, representado de forma holística na estrutura e em padrões de atividades de consideráveis regiões sobrepostas umas às outras, e não em células de memória discretas, dispostas ordenadamente.[149] Seria então impraticável estabelecer um mapeamento simples entre os neurônios de um e outro cérebro de maneira que os pensamentos pudessem passar facilmente de um cérebro ao outro. Para que os pensamentos de um cérebro sejam inteligíveis a outro, esses pensamentos precisam ser decompostos e transformados em símbolos de acordo com um protocolo compartilhado que permita que tal simbologia seja corretamente interpretada pelo cérebro receptor. Esse é o papel desempenhado pela linguagem.

Em princípio, podemos imaginar ser possível transferir todo o trabalho cognitivo de articulação e interpretação para uma interface que poderia, de alguma maneira, ler o estado neural do cérebro remetente e enviá-lo para o cérebro receptor com o uso de um padrão de ativação personalizado. Mas isso nos leva para o segundo problema referente à criação de um ciborgue. Mesmo se deixarmos de lado todos os (imensos) desafios tecnológicos por trás do processo de ler e escrever, simultaneamente e de forma confiável, informações contidas em aproximadamente 1 bilhão de neurônios individuais, fica claro que criar uma interface dessa natureza é provavelmente um problema IA-completo (ou seja, resolver essa questão seria equivalente a criar uma inteligência artificial geral). Essa interface precisaria ter um componente capaz de mapear (em tempo real) os padrões de ativação de um cérebro para os padrões de ativação semanticamente equivalentes de outro cérebro. A compreensão detalhada dos diversos níveis de processamento neural

exigida para a realização dessa tarefa seria muito semelhante àquela necessária para a criação de uma IA neuromórfica.

Apesar dessas reservas, a "ciborguização" como meio de obter melhorias cognitivas não deixa de ser promissora. Estudos impressionantes do hipocampo de ratos demonstraram a viabilidade de uma prótese neural capaz de aprimorar o desempenho dos animais em um teste simples de memória de curto prazo.[150] Em sua versão atual, o implante coleta dados provenientes de uma ou duas dúzias de eletrodos localizados em uma área ("CA3") do hipocampo e os projeta em um número similar de neurônios localizados em outra área ("CA1"). Um microprocessador é programado para fazer a discriminação entre dois padrões de ativação diferentes da primeira área (que correspondem a duas memórias diferentes, "alavanca direita" ou "alavanca esquerda") e aprender como esses padrões são projetados na segunda. Essa prótese pode não apenas restaurar funções cerebrais, caso alguma conexão neural natural entre as duas áreas tenha sido bloqueada, mas, ao enviar um símbolo claro de um padrão de memória particular para a segunda área, ela pode aprimorar o desempenho na tarefa de memorização para além do que um rato normalmente seria capaz. Apesar de tecnicamente excepcionais para os padrões atuais, esses estudos deixam em aberto uma série de perguntas difíceis: O quão bem essa abordagem funcionaria se aplicada a um número maior de memórias? Conforme o número de neurônios envolvidos aumentasse, seria possível controlar a explosão combinatória que ameaçaria tornar impraticável aprender o mapeamento correto? Há algum custo embutido nessa melhora de desempenho, como a redução da habilidade de fazer generalizações além dos estímulos utilizados no experimento, ou a redução na capacidade de "desaprender" as associações quando o ambiente sofre alterações? As pessoas testadas ainda teriam algum benefício se — ao contrário dos ratos — elas pudessem se valer de alguma memória externa, como papel e caneta? Quão mais difícil seria a aplicação de um método semelhante em outras áreas do cérebro? Enquanto a prótese utilizada atualmente se beneficia da estrutura relativamente simples de partes do hipocampo (servindo, basicamente, como uma ponte unidirecional entre as áreas CA3 e CA1), outras estruturas do córtex envolvem processos de retroalimentação bem mais complicados, o que aumentaria muito a complexidade do diagrama de ligações necessárias e, presumivelmente, tornaria mais difícil decifrar as funcionalidades de qualquer grupo de neurônios envolvidos no processo.

Uma esperança para a criação de um ciborgue é que o cérebro, caso recebesse um implante permanente de um dispositivo de comunicação com recursos externos, passaria, com o tempo, a *aprender* como fazer um mapeamento efetivo entre seus estados cognitivos internos e os dados recebidos do dispositivo ou aceitos pelo dispositivo. Desse modo, o implante em si não precisaria ser inteligente; ao contrário, o cérebro é que, inteligentemente, se adaptaria à interface, do mesmo modo que o cérebro de uma criança aprende gradualmente a interpretar os sinais oriundos de receptores como os olhos e os ouvidos.[151] Mas aqui, novamente, é preciso nos perguntarmos qual de fato seria o ganho efetivo. Suponhamos que a plasticidade do cérebro seja tamanha que ele venha a ser capaz de detectar

padrões a partir de uma nova série de dados projetados arbitrariamente em alguma parte de seu córtex através de uma interface cérebro-computador: por que não projetar essas mesmas informações na retina, na forma de padrões visuais, ou na cóclea, em forma de som? Essa alternativa de baixa tecnologia evitaria milhares de complicações e, em ambos os casos, o cérebro seria capaz de lançar mão de seus mecanismos de reconhecimento de padrões e plasticidade de modo a aprender a extrair sentido dessas informações.

Redes e organizações

Outro caminho concebível para a superinteligência é a melhoria gradual de redes e organizações que ligam as mentes humanas individuais umas às outras e a diversos artefatos e softwares. A ideia aqui não é afirmar que isso aprimoraria a capacidade intelectual de indivíduos a ponto de torná-los superinteligentes, mas que um sistema de diversos indivíduos conectados e organizados em rede poderia alcançar alguma forma de superinteligência — uma "superinteligência coletiva", conceito que será melhor elaborado no próximo capítulo.[152]

A humanidade progrediu enormemente em termos de inteligência coletiva ao longo do curso da história e da pré-história. Esse progresso derivou de diversas fontes: inovações em tecnologias de comunicação como a escrita e a imprensa e, mais do que qualquer outra coisa, a introdução da linguagem em si; o aumento no tamanho da população mundial e na densidade habitacional; diversos avanços em técnicas organizacionais e normas epistêmicas e uma acumulação gradual de capital institucional. De modo geral, a inteligência coletiva de um sistema é limitada pela habilidade das mentes que o compõem, pelo esforço necessário para transmitir informações relevantes entre seus membros e pelas muitas distorções e ineficiências que permeiam as organizações humanas. Se os custos de comunicação forem reduzidos (incluindo não apenas os custos com equipamentos, mas também a latência de resposta, tempo e atenção gastos e outros fatores), organizações cada vez maiores e mais densamente conectadas se tornariam viáveis. O mesmo poderia acontecer se fossem encontradas soluções para driblar algumas das deformações burocráticas que deturpam o mundo das grandes organizações — além das inúteis disputas por status, ambições exageradas, ocultação ou falsificação de informações e outros problemas. Mesmo soluções parciais para essas questões seriam capazes de gerar lucros significativos para a inteligência coletiva.

As inovações tecnológicas e institucionais que poderiam contribuir para o crescimento da nossa inteligência coletiva são numerosas e variadas. Por exemplo, mercados preditivos subsidiados poderiam implementar normas esclarecedoras e aprimorar as projeções em torno de questões científicas e sociais.[153] Detectores de mentiras (caso fosse possível produzir detectores confiáveis e fáceis de usar) poderiam reduzir o alcance das mentiras nas relações humanas.[154] Detectores capazes de identificar autoenganos poderiam ser ainda mais poderosos.[155] Mesmo sem o surgimento de tecnologias cerebrais modernas, alguns tipos de fraude podem se tornar cada vez mais difíceis de se cometer, graças aos diferentes

tipos de dados disponíveis, incluindo históricos e reputações individuais ou a promulgação de normas epistêmicas rígidas e de uma cultura de racionalidade. A vigilância voluntária e involuntária permitirá o acúmulo de uma enorme gama de informações a respeito do comportamento humano. Sites de redes sociais já são usados por mais de 1 bilhão de pessoas para o compartilhamento de informações pessoais: muito em breve, esses indivíduos poderão fazer uploads contínuos de gravações do seu dia a dia, realizadas através das câmeras e microfones de seus smartphones ou na armação de seus óculos. A análise automatizada desses dados possibilitará diversas novas aplicações (tanto malignas como benignas, é claro).[156]

O crescimento da inteligência coletiva também pode vir de melhorias econômicas e organizacionais e do aumento da parcela da população mundial que seja escolarizada e digitalmente conectada e integrada à cultural intelectual global.[157]

A internet se sobressai como uma fronteira particularmente dinâmica para inovações e experimentações. Muito de seu potencial se mantém ainda inexplorado. O contínuo desenvolvimento de uma web inteligente, com suporte maior para deliberações, redução da parcialidade e agregação de opiniões, pode contribuir consideravelmente para o incremento da inteligência coletiva da humanidade como um todo ou de algum grupo em particular.

Mas o que dizer a respeito da ideia, muito mais extravagante, de que a internet poderia um dia “acordar”? Será que a internet poderia se tornar mais do que apenas a espinha dorsal de uma superinteligência coletiva fragilmente integrada, vindo a ser algo como um crânio virtual que abriga um superintelecto emergente unificado? (Essa é uma das formas pelas quais a superinteligência poderia surgir de acordo com o influente artigo que Vernor Vinge escreveu em 1993, no qual o autor cunhou o termo “singularidade tecnológica”.)[158] Contrariamente a essa ideia, poderíamos argumentar que desenvolver uma inteligência de máquina já seria um feito extremamente difícil e demandaria um árduo trabalho de engenharia, sendo inconcebível supor que pudesse surgir de forma *espontânea*. Contudo, não necessariamente seria o caso de que alguma versão futura da internet, de repente e por acaso, pudesse se tornar superinteligente. Uma versão mais plausível desse cenário seria aquela em que a internet acumula avanços gerados a partir do trabalho de inúmeras pessoas ao longo de muitos anos — criando melhores ferramentas de busca e algoritmos de filtragem de informação, formatos mais eficazes de representação de dados, agentes de
software autônomos mais eficientes e protocolos mais efetivos de controle da interação entre esses diferentes agentes de software —, e essa miríade de avanços incrementais poderia, enfim, criar a base para uma forma mais unificada de inteligência na web. É até possível conceber que um sistema cognitivo baseado na web, supersaturado de capacidade computacional e de todos os outros recursos necessários para uma explosão de crescimento, com exceção de um elemento crucial, poderia então, uma vez que esse último elemento fosse adicionado à receita, se tornar superinteligente. Esse tipo de cenário, todavia, converge em outro caminho possível à superinteligência — aquele da inteligência artificial geral, já discutido anteriormente.

Resumo

A existência de diversos caminhos que podem nos levar à superinteligência deveria aumentar a confiança de que um dia a alcançaremos. Ainda que um desses caminhos seja eventualmente bloqueado, ainda nos restariam opções.

O fato de haver muitos caminhos não significa que existam múltiplos destinos. Mesmo que alcemos primeiro uma amplificação significativa da inteligência seguindo caminhos que não incluam as máquinas, isso não tornaria irrelevante a inteligência de máquina. Muito pelo contrário: um aumento da inteligência, quer tenha se originado por meio de melhorias biológicas ou organizacionais, aceleraria o desenvolvimento científico e tecnológico e, consequentemente, o surgimento de formas mais radicais de amplificação da inteligência, como a emulação completa do cérebro e a IA.

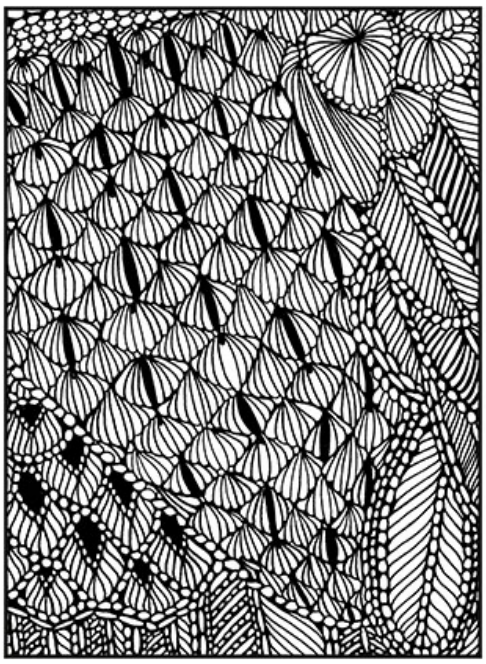
Isso não quer dizer que seja indiferente a maneira pela qual seria alcançada a superinteligência de máquina. O caminho utilizado para chegar até lá poderá fazer uma grande diferença no eventual resultado. Mesmo que as capacidades finais obtidas não dependam muito da trajetória tomada, a maneira como essas capacidades serão usadas — o grau de controle que os seres humanos terão sobre seus desejos e ordens — pode muito bem depender de nossa abordagem. Por exemplo, aprimoramentos na inteligência biológica ou organizacional podem aumentar nossa habilidade de antecipar riscos, conduzindo-nos ao desenvolvimento de uma superinteligência de máquina mais segura e benéfica. (Uma avaliação estratégica completa envolve uma série de complexidades e aguardará até o capítulo 14.)

Uma verdadeira superinteligência (diferente de incrementos marginais nos atuais níveis de inteligência) pode ser alcançada primeiro, muito possivelmente, pelo caminho da IA. Existem, contudo, diversas incertezas fundamentais ao longo desse caminho. Isso torna difícil uma avaliação precisa a respeito do tamanho desse caminho ou da quantidade de obstáculos que podem ser encontrados ao longo do percurso. A emulação completa do cérebro também tem chance de ser a rota mais rápida até a superinteligência. Uma vez que o progresso por esse caminho depende principalmente de avanços tecnológicos incrementais, e não de descobertas teóricas, existe uma boa possibilidade de que essa via seja bem-sucedida no futuro. Entretanto, ainda que o progresso de emulação completa do cérebro fosse rápido, muito provavelmente a inteligência artificial seria a primeira a cruzar a linha de chegada: isso se dá por conta da possibilidade de criação de IAs neuromórficas com base em emulações parciais.

Aperfeiçoamentos biológicos da cognição são claramente factíveis, particularmente aqueles baseados em seleção genética. O processo iterativo de seleção de embriões aparece, atualmente, como uma tecnologia especialmente promissora. Entretanto, se comparados aos possíveis avanços em inteligência de máquina, os aperfeiçoamentos biológicos seriam relativamente lentos e graduais. Eles resultariam, quando muito, em formas relativamente fracas de superinteligência (em breve, falarei mais sobre o assunto).

A clara viabilidade do aperfeiçoamento biológico deveria aumentar nossa confiança em relação ao surgimento futuro da inteligência de máquina, uma vez que cientistas e engenheiros biologicamente aperfeiçoados serão capazes de fazer progressos em maior quantidade e de forma mais rápida que seus pares *au naturel*. Especialmente num cenário em que a inteligência de máquina surgisse apenas depois de meados deste século, a entrada em cena de grupos de indivíduos mais aprimorados cognitivamente teria um papel cada vez maior em futuras descobertas.

As interfaces cérebro-computador parecem um caminho pouco provável até a superinteligência. Melhorias nas redes e organizações poderiam resultar em formas fracas de superinteligência coletiva no longo prazo; mas é mais provável que elas exerçam uma função inicial facilitadora semelhante àquela do aperfeiçoamento biológico, aumentando gradualmente a habilidade efetiva dos seres humanos para a resolução de problemas intelectuais. Em comparação ao aperfeiçoamento biológico, as melhorias nas redes e organizações farão diferença mais cedo — na verdade, tais avanços estão ocorrendo de maneira contínua e, atualmente, já exercem um impacto significativo. Contudo, as melhorias nas redes e organizações poderão trazer menos avanços em nossa capacidade de solução de problemas do que o aperfeiçoamento biológico da cognição — para antecipar uma distinção que será feita no próximo capítulo, elas aumentariam a “inteligência coletiva”, e não a “inteligência de qualidade”.



03

CAPÍTULO 3

FORMAS DE SUPERINTELIGÊNCIA

Afinal, o que exatamente queremos dizer com o termo “superinteligência”? Embora não desejemos nos atolar em lamaçais terminológicos, algo precisa ser dito para o esclarecimento dos conceitos aqui utilizados. Este capítulo identifica três formas diferentes de superinteligência e defende que, de acordo com o que é relevante na prática, elas são equivalentes. Também será demonstrado que o potencial para a inteligência em um substrato de máquina é muito maior do que em um substrato biológico. As máquinas possuem uma série de vantagens fundamentais capazes de lhes proporcionar uma superioridade incrível. Os humanos biológicos, mesmo se aperfeiçoados, serão ultrapassados.

Muitas máquinas e animais não humanos já atuam em níveis super-humanos em determinadas áreas. Morcegos interpretam sinais de radar melhor do que o homem, calculadoras já nos ultrapassaram em aritmética, e os programas de xadrez já são capazes de nos vencer. A gama de funções específicas que podem ser melhor executadas por softwares continuará a se expandir. Mas, ainda que os sistemas de processamento de informação especializados tenham muitas funções, há questões profundas que surgem somente com uma perspectiva na qual intelectos de máquina possuam uma inteligência geral capaz de substituir os humanos em todas as áreas de conhecimento.

Conforme afirmado anteriormente, usamos o termo “superinteligência” para fazer referência a intelectos que ultrapassam em muito as melhores mentes humanas em diversos aspectos cognitivos gerais. Mas tal conceito ainda é bastante vago, pois, sob essa definição, diferentes tipos de sistemas com atributos de desempenho um tanto distintos poderiam ser qualificados como superinteligências. Para avançar na análise, é útil desagregar essa noção simples de superinteligência distinguindo nela diferentes grupos de supercapacidades intelectuais. Há muitas maneiras pelas quais essa decomposição pode ser feita. Aqui, diferenciaremos três formas: superinteligência rápida, superinteligência coletiva e superinteligência de qualidade.

Superinteligência rápida

A superinteligência rápida é um intelecto exatamente igual à mente humana, porém mais rápido. Conceitualmente, é a forma de superinteligência mais fácil de ser analisada.[159] É possível definir a superinteligência rápida da seguinte forma:

> **Superinteligência rápida:** um sistema que pode fazer tudo o que um intelecto humano é capaz de fazer, porém muito mais rapidamente.

Por “muito”, queremos dizer algo em torno de “múltiplas ordens de magnitude”. Mas em vez de tentar eliminar qualquer imprecisão remanescente na definição, confiaremos que o leitor terá a sensatez suficiente para interpretá-la.[160]

O exemplo mais simples da superinteligência rápida seria uma emulação completa do cérebro executada em um hardware veloz.[161] Uma emulação operando com uma velocidade 10 mil vezes maior que a de um cérebro biológico seria capaz de ler um livro em alguns segundos e escrever uma tese de doutorado em uma tarde. Se a velocidade fosse 1 milhão de vezes maior, uma emulação poderia realizar o trabalho intelectual de um milênio em apenas um dia de trabalho.[162]

Para uma mente rápida como essa, os eventos no mundo externo parecem acontecer em câmera lenta. Imagine que sua mente opere 10 mil vezes mais rápido. Se o seu amigo de carne e osso derrubasse uma xícara, você poderia assisti-la por várias horas caindo vagarosamente em direção ao chão, como um cometa voando silenciosamente através do espaço em direção a um planeta distante; e conforme a antecipação da colisão se propaga tardiamente pela massa cinzenta de seu amigo e de lá em direção ao sistema nervoso periférico, você poderia observar o corpo dele assumindo gradualmente a expressão de surpresa — tempo suficiente para que você não apenas peça outra xícara, mas também leia uns dois artigos científicos e tire um cochilo.

Em virtude dessa expansão aparente de tempo do mundo material, a superinteligência rápida teria preferência por trabalhar com objetos digitais. Ela poderia viver em uma realidade virtual e operar na economia da informação. Outra opção seria interagir com o meio ambiente físico através de manipuladores que operem em nanoescala, uma vez que membros em escalas tão pequenas como essas poderiam operar mais rápido do que apêndices macroscópicos. (A frequência característica de um sistema tende a ser inversamente proporcional à sua escala de comprimento.)[163] Uma mente rápida poderia interagir preferencialmente com outras mentes rápidas em vez de se comunicar com as lerdas mentes humanas.

Conforme as mentes ficam mais rápidas, a velocidade da luz se torna um crescente e importante obstáculo, na medida em que mentes mais rápidas enfrentam custos de oportunidade maiores do uso do seu tempo para se deslocar ou se comunicar através de distâncias longas.[164] A luz é aproximadamente 1 milhão de vezes mais rápida que um avião a jato, logo, um agente digital com aceleração mental de aproximadamente 1.000.000× e um viajante humano contemporâneo precisariam da mesma quantidade de tempo subjetivo para viajar ao redor do mundo. Fazer uma ligação de longa distância para alguém ou comparecer “pessoalmente” levaria o mesmo tempo, embora a ligação demandasse menos custos, pois exigiria menos largura de banda. Agentes com grande aceleração mental que queiram conversar por um longo período podem achar vantajoso se deslocar para perto um do outro. Mentes extremamente rápidas com necessidade de interação frequente (tais como membros de um grupo de trabalho) poderiam residir em computadores localizados no mesmo edifício para evitar latências frustrantes.

Superinteligência coletiva

Outra forma de superinteligência é um sistema que pode alcançar um desempenho superior por meio da agregação de um grande número de inteligências menores.

> **Superinteligência coletiva:** um sistema composto de um grande número de intelectos menores, de forma que o seu desempenho total supere significativamente, em diversas áreas gerais do conhecimento, qualquer sistema cognitivo atual.

A superinteligência coletiva é conceitualmente menos clara do que a superinteligência rápida.[165] Porém, empiricamente ela é mais familiar. Embora não tenhamos experiência com mentes no nível humano que difiram significativamente em velocidade, temos vasta experiência com a inteligência coletiva, sistemas compostos por quantidades variadas de componentes de nível humano trabalhando em conjunto com graus de eficiência variados. Empresas, grupos de trabalho, redes de fofocas, grupos de defesa de direitos, comunidades acadêmicas, países e até mesmo a humanidade como um todo podem — se adotarmos uma perspectiva um tanto abstrata — ser vistos como "sistemas" vagamente definidos capazes de resolver diferentes categorias de problemas intelectuais. A partir de nossas experiências, temos alguma noção do quão facilmente algumas tarefas sucumbem às iniciativas de organizações de variados tamanhos e composições.

A inteligência coletiva se sobressai na resolução de problemas que podem ser facilmente divididos em partes, de modo que as soluções para esses subproblemas possam ser encontradas paralelamente e verificadas de maneira independente. Tarefas como a construção de uma nave espacial ou a administração de uma franquia de hambúrgueres oferecem uma miríade de oportunidades para a divisão do trabalho: engenheiros diferentes trabalham em componentes diferentes de uma nave espacial, empregados diferentes operam restaurantes diferentes. Em uma universidade, a rígida divisão de pesquisadores, alunos, periódicos, financiamentos e prêmios em disciplinas separadas e independentes — embora não favorável para o tipo de trabalho representado por este livro — pode (somente em um estado de espírito conciliatório e calmo) ser vista como uma acomodação necessária para permitir que um número maior de indivíduos e grupos com motivações diversas contribuam para o crescimento do conhecimento humano enquanto trabalham de forma relativamente independente, cada um arando sua própria horta.

A inteligência coletiva de um sistema pode ser aperfeiçoada através da expansão da quantidade ou da qualidade dos intelectos que a constituem, ou por meio da melhoria da qualidade de sua organização.[166] Para se obter uma *superinteligência* coletiva a partir de qualquer inteligência coletiva que exista atualmente, seria necessário um altíssimo grau de aperfeiçoamento. O sistema resultante precisaria ser capaz de superar significativamente qualquer inteligência coletiva atual ou outro sistema cognitivo em diversas áreas gerais de conhecimento. Um novo formato de conferência que permita que os estudiosos troquem informações de forma mais eficiente, ou um novo algoritmo de filtragem colaborativa de informações que melhor preveja as preferências dos consumidores de livros e filmes, não poderiam por si só

se aproximar de algo que se assemelhe a uma superinteligência coletiva. Isso não ocorreria com um aumento de 50% da população mundial, nem com uma melhora da metodologia pedagógica que permitisse que os alunos finalizassem um dia de estudo em quatro horas em vez de seis horas. Para que fossem preenchidos os critérios de uma superinteligência coletiva, seria necessário um crescimento muito mais intenso da capacidade cognitiva coletiva da humanidade.

Note que o patamar para uma superinteligência coletiva é indexado aos níveis de desempenho atuais — ou seja, o início do século XXI. Ao longo da pré-história humana e novamente no decorrer da história humana, a inteligência coletiva da humanidade cresceu significativamente. A população mundial, por exemplo, aumentou pelo menos mil vezes desde o Plistoceno.[167] Apenas com base nesse dado, poderíamos considerar que os níveis atuais de inteligência coletiva humana estariam se aproximando aos níveis de uma superinteligência se usássemos *como parâmetro o período Plistoceno*. Algumas melhorias nas tecnologias de comunicação — principalmente a linguagem falada, mas talvez também a formação de cidades, a escrita e a imprensa — podem, individualmente ou de forma combinada, ter concedido impulsos gigantescos. Nesse sentido, se nossa capacidade intelectual coletiva de resolução de problemas experimentasse outra inovação de impacto comparável, esse novo processo poderia resultar em uma superinteligência coletiva.[168]

A esta altura, um determinado tipo de leitor se sentiria tentado a insinuar que a sociedade moderna não parece ser assim tão inteligente. Talvez alguma decisão política não muito bem-vinda tenha acabado de ser tomada no país de origem do leitor e a aparente falta de sabedoria dessa decisão agora paira em sua mente como uma evidência da incapacidade mental da era moderna. E, afinal, não estaria a humanidade contemporânea idolatrando o consumismo, dilapidando os recursos naturais, poluindo o meio ambiente e dizimando a biodiversidade, ao mesmo tempo que falha em corrigir as gritantes injustiças globais e negligência importantes valores humanos ou espirituais? Todavia, deixando de lado a questão de como as limitações da modernidade se comparam aos fracassos não tão irrelevantes de épocas anteriores, nada na nossa definição de superinteligência coletiva implica que uma sociedade com maior inteligência coletiva se encontre necessariamente em melhores condições. A definição nem mesmo sugere que uma sociedade com uma maior inteligência coletiva seja *mais sábia*. Podemos pensar na sabedoria como a habilidade de entender aproximadamente aquilo que é importante. Então é possível imaginar uma organização composta por um amplo quadro de trabalhadores dotados de conhecimento que trabalham de forma eficiente e coordenada, os quais, coletivamente, podem resolver problemas intelectuais em áreas diferentes do conhecimento. Essa organização, suponhamos, pode operar diferentes tipos de negócios, inventar vários tipos de tecnologia e otimizar a maioria dos processos. Ainda assim, ela pode entender de forma totalmente errada algumas questões centrais de um cenário mais amplo — por exemplo, pode falhar em tomar precauções apropriadas contra riscos existenciais — e, como resultado, almejar um curto e explosivo surto de crescimento que terminasse de forma nada gloriosa em um colapso total. Tal organização poderia ter um grau muito alto de inteligência

coletiva e, caso esse grau fosse suficientemente alto, a organização seria uma superinteligência coletiva. Deveríamos resistir à tentação de juntar cada atributo normativamente desejável em um conceito amórfico gigante de funcionamento mental, como se ninguém pudesse achar uma característica admirável sem que as outras estejam igualmente presentes. Em vez disso, devemos reconhecer que podem existir sistemas de processamento de informação instrumentalmente poderosos — sistemas inteligentes — que não são nem inerentemente bons, nem confiavelmente sábios. Mas esse assunto será revisitado no capítulo 7.

A superinteligência coletiva poderia ser fracamente ou fortemente integrada. Para ilustrar um caso de uma superinteligência coletiva fracamente integrada, imagine um planeta chamado *MegaTerra*, que possui o mesmo nível de tecnologias de comunicação e coordenação que temos atualmente no planeta Terra real, porém com uma população 1 milhão de vezes maior. Com uma população tão grande assim, a mão de obra intelectual no planeta MegaTerra seria correspondentemente maior do que a do nosso planeta. Suponha que exista ao menos um gênio cientista do calibre de Newton ou de Einstein para cada 10 bilhões de pessoas: então na MegaTerra haveria 700 mil gênios desse tipo vivendo contemporaneamente na companhia de vastas multidões de gênios ligeiramente menos talentosos. Novas ideias e tecnologias seriam desenvolvidas em um ritmo incrível, e a civilização global na MegaTerra constituiria uma superinteligência coletiva fracamente integrada.[169]

Se aumentássemos gradualmente o nível de integração de uma inteligência coletiva, ela poderia finalmente se tornar um *intelecto* unificado — uma única e grande “mente”, em vez de um simples conjunto de mentes humanas menores interagindo pouco umas com as outras.[170] Os habitantes da MegaTerra poderiam dar passos em direção a um intelecto unificado por meio do aprimoramento das tecnologias de comunicação e coordenação e com o desenvolvimento de maneiras mais eficientes de trabalho em conjunto para que muitos indivíduos possam trabalhar ao mesmo tempo em qualquer problema intelectual complexo. Uma superinteligência coletiva poderia, então, após obter um ganho suficiente em integração, se tornar uma
“superinteligência de qualidade”.

Superinteligência de qualidade

Podemos distinguir uma terceira forma de superinteligência.

> **Superinteligência de qualidade:** *um sistema que é no mínimo tão rápido quanto uma mente humana e qualitativamente muito mais inteligente.*

Assim como a inteligência coletiva, a inteligência de qualidade também é um conceito ligeiramente nebuloso; e nesse caso a dificuldade é agravada pela falta de experiência com qualquer variação na qualidade da inteligência que ultrapasse o limite superior da presente distribuição humana. É possível, entretanto, ter uma pequena noção se considerarmos alguns casos.

Primeiro, nós podemos expandir o alcance dos nossos pontos de referência considerando animais não humanos dotados de uma inteligência de qualidade inferior. (Essa observação não deve, de forma alguma, ser encarada como especismo. Um peixe-zebra possui uma qualidade de inteligência perfeitamente adaptada às suas necessidades ecológicas; mas a perspectiva relevante aqui é mais antropocêntrica: estamos focados no desempenho de tarefas cognitivas complexas e *humanamente* relevantes.) Animais não humanos carecem de uma linguagem estruturada complexa; a maioria não é capaz de construir e usar ferramentas enquanto outros se utilizam apenas de instrumentos rudimentares; têm capacidade extremamente restrita de fazer planos a longo prazo; e possuem habilidade muito limitada de raciocínio abstrato. Tais limitações não são completamente explicadas pela falta de velocidade ou de inteligência coletiva nas mentes de animais não humanos. Em termos de poder computacional bruto, os cérebros humanos são, provavelmente, inferiores aos dos animais de grande porte, incluindo os elefantes e as baleias. E embora a complexa civilização tecnológica da humanidade fosse impossível sem nossa vantagem massiva em inteligência coletiva, nem todas as capacidades cognitivas especificamente humanas dependem da inteligência coletiva. Muitas são, inclusive, altamente desenvolvidas mesmo em pequenos grupos isolados de caçadores-coletores.[171] E muitas não estão tão desenvolvidas em alguns grupos altamente organizados de animais não humanos, como chimpanzés e golfinhos treinados intensamente por instrutores humanos, ou formigas que vivem nas suas próprias sociedades amplas e organizadas. Evidentemente, as notáveis conquistas intelectuais do *Homo sapiens* são, em grande medida, atribuíveis a características específicas da arquitetura do nosso cérebro, as quais dependem de uma herança genética única que não é compartilhada por outros animais. Essa observação pode nos ajudar a ilustrar o conceito da superinteligência de qualidade: é uma inteligência de qualidade no mínimo superior à inteligência humana, assim como a qualidade da inteligência humana é superior à dos elefantes, golfinhos e chimpanzés.

Também é possível ilustrar o conceito da superinteligência de qualidade destacando os déficits cognitivos em áreas específicas que podem afligir os indivíduos humanos, especialmente déficits que não são causados por demência generalizada ou outras condições associadas à destruição total dos recursos neurocomputacionais do cérebro. Considere, por exemplo, indivíduos com transtorno do espectro autista que tenham déficits cognitivos sociais, embora apresentem bom desempenho em outras áreas cognitivas; ou indivíduos com amusia congênita, que não são capazes de cantarolar ou reconhecer sons simples, mas possuem desempenho normal em outros aspectos. Muitos outros exemplos poderiam ser retirados da literatura de neuropsiquiatria, que é repleta de estudos de caso de pacientes com déficits restritos causados por anormalidades genéticas ou traumas cerebrais. Tais exemplos demonstram que os adultos humanos normais possuem uma série de talentos cognitivos notáveis não apenas pelo fato de apresentarem uma quantidade suficiente de poder de neuroprocessamento geral ou mesmo uma quantidade suficiente de inteligência geral: circuitos neurais especializados também são necessários. Essa observação sugere a ideia de

talentos cognitivos possíveis, mas não realizados, talentos que nenhum humano real possui, embora outros sistemas inteligentes — com poder computacional comparável ao do cérebro humano — dotados de tais talentos pudessem ter enormes vantagens nas suas habilidades de realizar uma ampla gama de tarefas estrategicamente relevantes.

Então, considerando animais não humanos e indivíduos humanos com déficits cognitivos em áreas específicas, podemos elaborar alguma noção de diferentes qualidades de inteligência e a real diferença prática existente entre elas. Se tivessem faltado ao *Homo sapiens*, por exemplo, os módulos cognitivos que possibilitam representações linguísticas complexas, a espécie poderia ter sido apenas mais um tipo de símio vivendo em harmonia com a natureza. Por outro lado, se nos fosse *dado* um novo conjunto de módulos que nos fornecesse uma vantagem comparável àquela de ser capaz de formar representações linguísticas complexas, nós teríamos nos tornado superinteligentes.

Alcance direto e indireto

A superinteligência em qualquer uma dessas formas poderia, ao longo do tempo, desenvolver a tecnologia necessária para criar qualquer uma das outras. Os *alcances indiretos* dessas três formas de superinteligência são, portanto, iguais. Nesse sentido, o alcance indireto da inteligência humana atual está também na mesma classe de equivalência, supondo que sejamos capazes, em algum momento futuro, de criar alguma forma de superinteligência. Mesmo assim, há um sentido em que as três formas de superinteligência ficam muito mais próximas umas das outras: qualquer uma delas poderia criar outras formas de superinteligência mais rapidamente do que nós seríamos capazes, considerando nosso ponto de partida atual.

Os *alcances diretos* dessas três formas diferentes de superinteligência são mais difíceis de comparar. Pode não haver uma ordem definida. Suas respectivas capacidades dependem do grau em que esses alcances instanciam suas respectivas vantagens — *quão* rápida é a velocidade de uma superinteligência rápida, *quão* qualitativamente superior é uma superinteligência de qualidade, e assim por diante. No máximo, podemos dizer que, *ceteris paribus*, a superinteligência rápida se sobressai em tarefas que exigem a rápida execução de uma longa série de passos que devem ser realizados sequencialmente, enquanto a superinteligência coletiva se sobressai em tarefas que admitam uma decomposição analítica em subtarefas paralelizáveis e tarefas que demandem a combinação de várias perspectivas diferentes e conjuntos de aptidões. Em algum sentido vago, a superinteligência de qualidade seria a mais capaz de todas as formas, na medida em que fosse capaz de perceber e resolver problemas que estão, para todos os propósitos práticos, além do alcance *direto* da superinteligência rápida e da superinteligência coletiva.[172]

Em algumas áreas, a quantidade é um substituto ruim para a qualidade. Um gênio solitário trabalhando em um quarto revestido de cortiça pode escrever *Em busca do tempo perdido*. Uma obra-prima equivalente poderia ser produzida se

recrutássemos um edifício comercial repleto de escritores medíocres?[173] Mesmo dentro do espectro da variação humana atual, podemos observar que algumas funções se beneficiam muito mais do trabalho de uma mente brilhante do que de esforços conjuntos de uma multidão de mentes medíocres. Se aumentarmos nossa perspectiva de análise de modo a incluir mentes *superinteligentes*, teremos que encarar a possibilidade da existência de problemas intelectuais solucionáveis somente por uma superinteligência, os quais jamais poderiam ser solucionados por qualquer coletividade, por maior que ela seja, de humanos não aperfeiçoados.

Portanto, devem existir alguns problemas que são solucionáveis por uma superinteligência de qualidade e talvez por uma superinteligência rápida, mas que não podem ser solucionados por uma superinteligência coletiva fracamente integrada (a não ser que ela amplificasse primeiro sua própria inteligência).[174] Nós não podemos ver claramente quais são esses problemas, mas podemos caracterizá-los em termos gerais.[175] Tendencialmente, seriam problemas que envolveriam múltiplas interdependências complexas, as quais não permitiriam soluções em etapas que possam ser verificadas de maneira independente: problemas que, portanto, não podem ser resolvidos de maneira fragmentada e que poderiam exigir maneiras qualitativamente novas de entendimento ou novas estruturas representacionais que sejam muito profundas ou muito complicadas para serem descobertas ou efetivamente utilizadas pelos atuais mortais. Alguns tipos de criações artísticas e raciocínios estratégicos podem entrar nessa categoria, assim como alguns tipos de descobertas científicas. Há quem especule que o atraso e as oscilações do progresso da humanidade em muitos dos “problemas eternos” da filosofia são devidos à inadequação do córtex humano para trabalhos filosóficos. Nessa ótica, nossos filósofos mais célebres são como cachorros andando com suas patas posteriores — mal atingindo o nível mínimo de desempenho necessário para realizar suas atividades.[176]

Vantagens da inteligência digital

Pequenas alterações no volume e conexões do cérebro podem ter grandes consequências, conforme podemos observar quando comparamos as conquistas intelectuais e tecnológicas dos humanos com as dos outros primatas. As mudanças muito maiores em recursos computacionais e arquitetura que serão possibilitadas pela inteligência de máquina provavelmente terão consequências ainda mais profundas. Para nós é difícil, talvez impossível, intuir a respeito das aptidões de uma superinteligência; mas pelo menos conseguimos ter um indício dessas possibilidades se olharmos para algumas das vantagens disponíveis para as mentes digitais. As vantagens do hardware são mais fáceis de apreciar:

- *Velocidade dos elementos computacionais.* Neurônios biológicos operam a uma velocidade máxima de 200 Hz, sendo sete ordens completas de magnitude mais lentos do que um microprocessador moderno (~2 GHz).[177] Como consequência, o cérebro humano é forçado a depender de uma paralelização maciça e é incapaz de realizar rapidamente qualquer cálculo que

exija um grande número de operações sequenciais.[178] (O cérebro não pode executar mais do que cem operações sequenciais — talvez apenas algumas dezenas — para realizar qualquer atividade que leve menos de um segundo.) Porém, muitos dos algoritmos mais relevantes para a área de programação e ciência da computação não são facilmente paralelizáveis. Muitas tarefas cognitivas poderiam ser executadas de forma mais eficiente se o suporte nativo do cérebro para algoritmos paralelizáveis de reconhecimento de padrões fosse complementado e integrado com um suporte para processamento sequencial rápido.

• *Velocidade interna de comunicação.* Axônios possuem potenciais de ação a velocidades de 120 m/s ou menos, enquanto núcleos de processamento eletrônico podem se comunicar opticamente na velocidade da luz (300.000.000 m/s).[179] A lentidão dos sinais neurais limita o quão grande um cérebro biológico pode ser enquanto funciona como uma unidade única de processamento. Por exemplo, para realizar uma viagem de ida e volta entre dois elementos em um sistema com latência de menos de 10 ms, cérebros biológicos precisam ser menores do que 0,11 m³. Um sistema eletrônico, por outro lado, poderia ter $6{,}1\times10^{17}$ m³, aproximadamente o tamanho de um planeta anão: dezoito ordens de magnitude maior.[180]

• *Números de elementos computacionais.* O cérebro humano possui um pouco menos de 100 bilhões de neurônios[181] e é aproximadamente três vezes e meia maior que o cérebro dos chimpanzés (embora tenha apenas um quinto do tamanho do cérebro de uma baleia cachalote).[182] O número de neurônios em uma criatura biológica é claramente limitado pelo volume do crânio e pela capacidade metabólica, mas outros fatores também podem ser significativos para cérebros maiores (tais como controle de temperatura, tempo de desenvolvimento e atrasos de transmissão de sinal — ver item anterior.) Em contrapartida, o hardware de um computador pode ter seus limites físicos expandidos indefinidamente a níveis altíssimos.[183] Supercomputadores podem ser do tamanho de um galpão industrial ou até maiores, com capacidade remota suplementar adicionada por meio de cabos de alta velocidade.[184]

• *Capacidade de armazenamento.* A memória de trabalho humana é capaz de armazenar apenas quatro ou cinco elementos de informação por vez.[185] Embora seja um equívoco comparar diretamente o tamanho de uma memória humana com a quantidade de RAM presente em um computador digital, é evidente que as vantagens de hardware das inteligências digitais tornarão possíveis memórias de trabalho maiores. Isso pode habilitar essas mentes a compreender intuitivamente relacionamentos complexos que os humanos só conseguem captar com dificuldade através de raciocínios lentos e penosos.[186] A memória humana de longo prazo também é limitada, embora não seja claro se somos capazes de exaurir sua capacidade de armazenamento ao longo do período de uma vida normal — o ritmo no qual acumulamos informação é muito lento. (De acordo com uma estimativa, o cérebro humano adulto armazena

aproximadamente 1 bilhão de bits — algumas ordens de magnitude abaixo da capacidade de um smartphone simples.)[187] Tanto a quantidade de informação armazenada quanto a velocidade com a qual ela pode ser acessada poderiam, portanto, ser bem maiores em um cérebro artificial do que em um cérebro biológico.

• *Confiabilidade, duração, sensores etc.* Inteligências de máquina podem ter várias outras vantagens de hardware. Por exemplo, neurônios biológicos são menos confiáveis que os transistores.[188] Como a computação com ruído necessita de esquemas de codificação redundantes que usam múltiplos elementos para codificar um único bit de informação, um cérebro digital pode ter alguns ganhos de eficiência com o uso de elementos computacionais confiáveis de alta precisão. Os cérebros se cansam depois de algumas horas de trabalho e começam a se degradar de forma permanente após algumas décadas de tempo subjetivo; microprocessadores não estão sujeitos a tais limitações. O fluxo de dados de entrada em uma inteligência de máquina poderia ser aumentado com a adição de milhões de sensores. Dependendo da tecnologia utilizada, a máquina poderia ter um hardware reconfigurável que pudesse ser otimizado de acordo com as necessidades exigidas por diferentes tarefas, enquanto a maior parte da arquitetura do cérebro é definida no nascimento ou só pode ser modificada lentamente (embora os detalhes de conectividade sináptica possam mudar em períodos curtos de tempo, como dias).[189]

Atualmente, o cérebro biológico ainda se sai melhor quando comparado aos computadores digitais em termos de poder computacional, embora supercomputadores de ponta estejam atingindo níveis de performance que se assemelham a uma estimativa plausível do que seria o poder de processamento do cérebro.[190] Mas os hardwares estão se aperfeiçoando rapidamente e seus limites máximos de desempenho são muito maiores do que aqueles encontrados nos substratos computacionais biológicos.

Mentes digitais também se beneficiarão de vantagens significativas de software:

• *Editabilidade.* É mais fácil experimentar variações de parâmetros em softwares do que em *wetware* neural. Por exemplo, com uma emulação completa do cérebro, é possível facilmente verificar quais seriam os efeitos de uma adição de neurônios em uma área cortical específica ou resultado do aumento ou diminuição da excitabilidade neuronal. Fazer tais experimentos em cérebros biológicos vivos seria bem mais difícil.

• *Duplicabilidade.* Com software, é possível criar, de forma rápida e arbitrária, muitas cópias de alta fidelidade para ocupar a base de hardware disponível. Cérebros biológicos, ao contrário, só podem ser reproduzidos lentamente, e cada novo exemplar inicia em um estado impotente, não lembrando nada do que seus pais aprenderam ao longo de suas vidas.

• *Coordenação de objetivos.* Coletividades humanas são repletas de ineficiências decorrentes do fato de ser praticamente impossível alcançar

uniformidade de propósito entre os membros de um grande grupo — pelo menos até que seja viável induzir submissão em larga escala por meio de drogas ou seleção genética. Um "clã de cópias" (um grupo de programas idênticos ou quase idênticos compartilhando o mesmo objetivo) evitaria tais problemas de coordenação.

• *Compartilhamento de memória.* Cérebros biológicos precisam de longos períodos de treinamento e aconselhamento enquanto mentes digitais poderiam adquirir novas memórias e habilidades através da troca de arquivos de dados. Uma população de 1 bilhão de cópias de um programa de IA poderia sincronizar suas bases de dados periodicamente, de forma que qualquer instância do programa saberia tudo o que as outras instâncias teriam aprendido na hora anterior. (Transferência direta de memórias exige formatos representacionais padronizados. Portanto, uma troca simples de conteúdo cognitivo de alto nível não seria possível entre duas inteligências de máquina de qualquer tipo e seria particularmente impossível nas primeiras emulações completas do cérebro.)

• *Novos módulos, modalidades e algoritmos.* A percepção visual nos parece algo fácil e simples, bem diferente da solução de problemas de geometria propostos em livros didáticos — isso acontece, apesar do fato de ser necessária uma enorme quantidade de cálculos para reconstruir, a partir de padrões de estimulação bidimensionais presentes em nossa retina, uma representação tridimensional de um mundo habitado por objetos reconhecíveis. A razão pela qual esse processo parece fácil se deve ao fato de possuirmos mecanismos neurais de baixo nível dedicados ao processamento de informações visuais. Esse processamento de baixo nível ocorre inconscientemente e de forma automática, sem drenar nossa energia mental ou atenção consciente. Percepção musical, uso da linguagem, cognição social e outras formas de processamento de informações que são "naturais" para nós, humanos, parecem ser igualmente apoiadas por módulos neurocomputacionais dedicados. Uma mente artificial que contasse com esse suporte especializado em outras áreas de cognição que se tornaram importantes no mundo contemporâneo — como engenharia, programação de computadores e estratégia de negócios — teria grandes vantagens em relação a mentes como as nossas, que dependem de uma cognição desorganizada e pouco específica para pensar a respeito de tais assuntos. Novos algoritmos também podem ser desenvolvidos para tirar vantagem das possibilidades exclusivas do hardware digital, como seu suporte para o processamento serial rápido.

As vantagens que *potencialmente* podem ser obtidas com a inteligência de máquina, hardware e software combinados, são gigantescas.[191] Mas quão rapidamente essas possíveis vantagens poderiam ser materializadas? Essa é a questão à qual nos dedicaremos agora.

04

CAPÍTULO 4

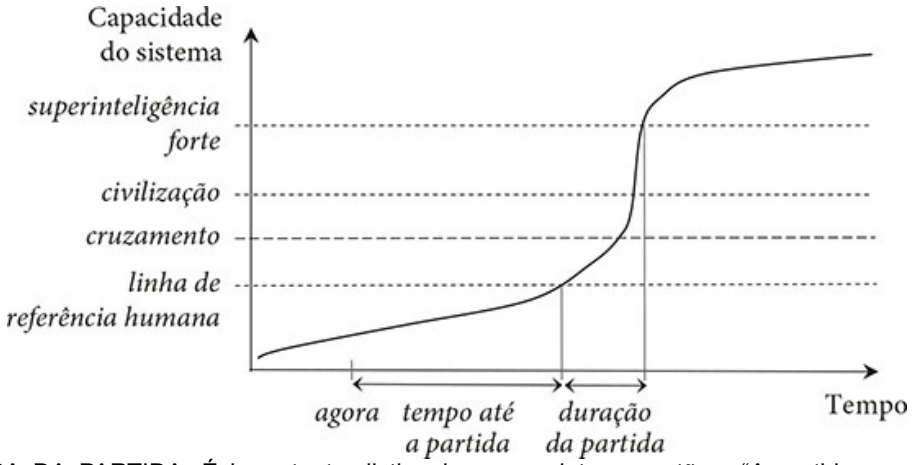
A CINÉTICA DE UMA EXPLOSÃO DE INTELIGÊNCIA

Uma vez que as máquinas tenham alcançado alguma forma de equivalência humana na habilidade de raciocínio, quanto tempo levaria, então, até que elas atingissem uma superinteligência radical? Essa seria uma transição lenta, gradual e prolongada? Ou seria repentina e explosiva? Este capítulo analisa a cinética de transição para a superinteligência como uma função do poder de otimização e da resistência do sistema. Consideramos o que sabemos ou o que podemos razoavelmente supor a respeito do comportamento desses dois fatores em relação à inteligência geral de nível humano.

Momento e velocidade da partida

Considerando que as máquinas, *em algum momento no futuro*, ultrapassarão consideravelmente a biologia em termos de inteligência geral, mas que *atualmente* a cognição das máquinas é muito mais limitada que a cognição humana, é possível se questionar a respeito de quão rapidamente essa ultrapassagem acontecerá. A questão que levantamos aqui deve ser claramente distinguida da questão que consideramos no capítulo 1, sobre o quão longe estamos, atualmente, de desenvolver uma máquina com nível humano de inteligência geral. Aqui a questão é: *se e quando tal máquina for desenvolvida, quanto tempo levará até que ela se torne radicalmente superinteligente?* Note que algumas pessoas podem achar que levará muito tempo até que as máquinas alcancem o nível humano, ou simplesmente podem não ter opinião formada sobre quanto tempo isso levará, e ainda ter uma opinião firme sobre o fato de que quando isso ocorrer a subsequente escalada em direção a uma superinteligência forte será muito rápida.

FIGURA 7



A ESTRUTURA DA PARTIDA. É importante distinguir as seguintes questões: “A partida ocorrerá? Se sim, quando?” e “Se e quando a partida ocorrer, o quão veloz ela seria?”. Seria possível, por exemplo, acreditar que

levará muito tempo para que a partida ocorra, mas que, uma vez que isso aconteça, o processo se dará rapidamente. Outra questão relevante (não ilustrada nesta figura) é: “Quão grande seria a parcela da economia mundial envolvida nesse processo?”. Essas perguntas são relacionadas, embora distintas.

Pode ser útil pensar a respeito desses assuntos esquematicamente, embora isso envolva ignorar temporariamente algumas ressalvas e detalhes complicadores. Considere, então, um diagrama que trace a capacidade intelectual do sistema mais avançado de inteligência de máquina como uma função do tempo (figura 7).

A linha horizontal, identificada como “linha de referência humana”, representa as capacidades intelectuais efetivas de um humano adulto representativo com acesso a fontes de informação e apoio tecnológico atualmente disponíveis em países desenvolvidos. No presente, o sistema mais avançado de IA está bem abaixo da linha de referência humana em qualquer medição aceitável de habilidade intelectual geral. Em algum momento no futuro, uma máquina poderá alcançar uma paridade aproximada com essa linha de referência (que estamos supondo ser fixa, por exemplo, no ano de 2014, ainda que as capacidades humanas individuais aumentem nos anos seguintes): isso marcaria o início da partida. As capacidades do sistema continuariam a crescer, e mais adiante o sistema alcançaria a paridade com a capacidade intelectual combinada de toda a humanidade (novamente baseada no presente): o que poderíamos chamar de “linha de referência da civilização”. Finalmente, se as habilidades do sistema continuarem a aumentar, ele atingirá uma “superinteligência forte” — um nível de inteligência muito maior do que a combinação de todos os recursos intelectuais da humanidade contemporânea. A conquista de uma superinteligência forte marca o final desse processo de partida, embora o sistema possa continuar a ganhar capacidade posteriormente. Em algum momento durante a fase de partida, o sistema poderá ultrapassar um ponto de referência que podemos chamar de “cruzamento”, um ponto no qual o aprimoramento do sistema se dá principalmente pelas ações do próprio sistema, e não em resposta a ações externas.[192] (A possível existência de tal “cruzamento” se tornará importante mais adiante neste capítulo, na subseção sobre poder de otimização e explosividade.)

Com essa imagem em mente, podemos diferenciar três classes de cenários de transição — cenários nos quais os sistemas progridem de um nível de inteligência humana para o de superinteligência —, tendo suas velocidades como base; ou seja, se eles representam uma partida lenta, rápida ou moderada.

Partida lenta

Uma partida lenta é aquela que ocorre em um intervalo de tempo longo, como décadas ou séculos. Cenários de partida lentos oferecem excelentes oportunidades para os processos políticos humanos se adaptarem e reagirem. Abordagens diferentes podem ser experimentadas e testadas em sequência. Novos especialistas podem ser treinados e credenciados. Campanhas populares podem ser mobilizadas pelos grupos que acreditem estar em desvantagem diante dos últimos desenvolvimentos. Se novos tipos de infraestrutura segura ou vigilância em massa dos pesquisadores de IA se mostrarem necessários, tais sistemas poderiam ser desenvolvidos e

implantados. Países que temessem uma corrida armamentista da IA teriam tempo para negociar novos tratados e desenvolver novos mecanismos de intervenção. A maioria dos preparativos realizados antes do início de uma partida lenta seria considerada obsoleta na medida em que melhores soluções gradualmente se tornariam aparentes à luz de uma nova era.

Partida rápida
Uma partida rápida é aquela que ocorre em um intervalo de tempo curto, como minutos, horas ou dias. Cenários de partida rápida oferecem poucas oportunidades para deliberações por parte dos seres humanos. Ninguém sequer notaria algo de diferente antes que o jogo já estivesse perdido. Em um cenário de partida rápida, o destino da humanidade dependeria essencialmente de preparações prévias. Nas versões mais lentas do cenário de partida rápida, algumas ações humanas simples seriam possíveis, semelhantes à abertura de uma "maleta nuclear"; mas qualquer uma dessas ações seria elementar ou teria sido planejada e pré-programada.

Partida moderada
Uma partida moderada é aquela que ocorre em um intervalo intermediário de tempo entre o curto e o longo, como meses ou anos. Esse cenário propicia aos humanos alguma chance de resposta, mas não muito tempo para analisar a situação, testar diferentes abordagens ou resolver problemas complicados de coordenação. Não há tempo suficiente para desenvolver e lançar novos sistemas (por exemplo, sistemas políticos, regimes de vigilância ou protocolos de segurança de rede), mas sistemas existentes poderiam ser aplicados aos novos desafios.

Durante uma partida lenta, haveria tempo suficiente para que as notícias se espalhassem. Em uma partida moderada, ao contrário, é possível que os acontecimentos fossem mantidos em sigilo enquanto se desenrolassem. O conhecimento poderia estar restrito a um pequeno grupo como em um programa secreto de pesquisa militar patrocinado pelo Estado. Projetos comerciais, pequenos grupos acadêmicos e alguns hackers trabalhando secretamente poderiam ser clandestinos — entretanto, se a possibilidade de uma explosão de inteligência estivesse no "radar" das agências nacionais de inteligência, como uma prioridade de segurança nacional, então os projetos privados mais promissores teriam uma boa chance de estar sob vigilância. O Estado onde o projeto estivesse sendo desenvolvido (ou uma potência dominadora estrangeira) teria, então, a opção de estatizar ou encerrar qualquer programa que apresentasse sinais de ter chegado perto de um ponto de partida. Partidas rápidas aconteceriam tão depressa que não haveria tempo para que a notícia se espalhasse ou para que qualquer pessoa pudesse preparar uma reação significativa. Porém, um indivíduo de fora poderia interferir *antes* do início da partida se ele acreditasse que um projeto em particular estivesse próximo de obter sucesso.

Cenários de partida moderada poderiam levar a uma turbulência geopolítica, social e econômica com indivíduos e grupos competindo para lucrar com o desenrolar das transformações. Caso ocorresse, essa agitação poderia impedir esforços para orquestrar uma resposta mais bem elaborada, ou conduziria a soluções mais radicais do que as permitidas por circunstâncias. Por exemplo, em um cenário de partida moderada onde emulações baratas e eficientes ou outras mentes digitais fossem gradualmente inundando o mercado de trabalho por um período de anos, poderíamos imaginar protestos em massa de trabalhadores demitidos que pressionariam os governos a aumentar o seguro-desemprego, implantar uma renda mínima para todos os cidadãos, arrecadar impostos especiais ou, ainda, impor aos empregadores que utilizassem trabalhadores emulados a exigência do pagamento de um salário-mínimo. Para que qualquer benefício oferecido por essas medidas não fosse apenas passageiro, o apoio a elas teria que estar garantido, de alguma forma, por estruturas permanentes de poder. Questões similares podem surgir caso a partida seja lenta em vez de moderada, mas o desequilíbrio e as mudanças rápidas de uma partida moderada poderiam apresentar oportunidades especialmente favoráveis para que pequenos grupos exercessem influência desproporcional.

Pode parecer, para alguns leitores, que, entre os três tipos de cenários, a partida lenta seja a mais provável, a moderada seja menos provável e a rápida seja totalmente impossível. É possível julgar fantasiosa a suposição de que o mundo pudesse ser rapidamente transformado e que, no curso de uma ou duas horas, a humanidade fosse deposta de sua posição de protagonista. Uma mudança tão rápida jamais ocorreu na história da humanidade, e os exemplos mais próximos — as revoluções agrícola e industrial — se desenrolaram em uma escala de tempo muito mais longa (de séculos a milênios no primeiro caso e de décadas a séculos no segundo caso). Dessa forma, a referência para o tipo de transição provocada por um cenário de partida rápida ou moderada, em termos de velocidade e magnitude das mudanças postuladas, é nula: não há nenhum precedente fora do mito ou da religião.[193]

Entretanto, este capítulo apresentará algumas razões para mostrar que um cenário de transição lenta é improvável. Se e quando uma partida ocorrer, tudo indica que ela será explosiva.

Para começar a analisar o quão rápida será a partida, podemos imaginar a taxa de aumento da inteligência de um sistema como uma função (monotonicamente crescente) de duas variáveis: a quantidade de “poder de otimização”, ou o esforço necessário para a criação de um projeto com uma certa qualidade, que é aplicado para aumentar a inteligência do sistema, e a capacidade de resposta do sistema à aplicação de uma dada quantidade de poder de otimização. Podemos denominar o inverso da capacidade de resposta como “resistência” e escrever:

$$\text{Taxa de mudança da inteligência} = \frac{\text{Poder de otimização}}{\text{Resistência}}$$

Na falta de alguma especificação de como quantificar inteligência, esforço de projeto e resistência, essa expressão é meramente qualitativa. Mas podemos ao

menos observar que a inteligência de um sistema aumentará rapidamente *caso* um grande esforço qualificado for aplicado à tarefa de aumentar sua inteligência e ela não seja tão difícil de ser aumentada *ou* haja um esforço de projeto não trivial e a resistência do sistema seja baixa (ou ambos). Se fosse possível estimar o esforço de projeto necessário para melhorar um determinado sistema, além do grau de melhorias que tal esforço produziria, poderíamos calcular a resistência do sistema.

Além disso, podemos observar que a quantidade de poder de otimização dedicada para melhorar o desempenho do sistema em algum aspecto varia ao longo do tempo e entre sistemas. A resistência de um sistema também pode variar dependendo do quanto o sistema já tenha sido otimizado. Geralmente as melhorias mais fáceis são realizadas primeiro, resultando em retornos reduzidos (resistência maior) assim que os casos mais fáceis se esgotem. Entretanto, também pode haver melhorias que tornem incrementos futuros mais fáceis, conduzindo a uma sequência de melhorias. O processo de montar um quebra-cabeça é iniciado de maneira simples — é fácil encontrar os cantos e as extremidades. Mas a resistência aumenta assim que as peças subsequentes se tornam mais difíceis de ser encaixadas. Mas assim que o quebra-cabeça está perto de ser finalizado, o número de peças soltas diminui e o processo fica fácil novamente.

Para procedermos com a nossa investigação devemos, portanto, analisar de que maneira a resistência e o poder de otimização podem variar em períodos críticos de tempo durante a partida. Nos ocuparemos dessa questão nas próximas páginas.

Resistência

Iniciemos com a discussão sobre a resistência. O panorama aqui depende do tipo do sistema que está sendo considerado. Para obter uma visão completa, primeiro consideraremos a resistência que seria encontrada ao longo dos caminhos para a superinteligência que não envolvam inteligência de máquina avançada. Veremos que a resistência ao longo desses caminhos parece ser relativamente alta. Feito isso, voltaremos para o caso principal, que é aquele no qual a partida envolve inteligência de máquina; e lá observaremos que a resistência em seu ponto crítico parece ser baixa.

Caminhos de inteligência de não máquinas

Aprimoramentos cognitivos por meio de melhorias na saúde pública e na dieta têm fornecido retornos cada vez menores.[194] Os maiores ganhos resultaram da eliminação de deficiências nutricionais severas, que já foram praticamente eliminadas, exceto nos países mais pobres. Quando uma dieta adequada é aumentada, há apenas um ganho de peso extra. Atualmente, a educação também está provavelmente sujeita a gerar retornos menores. A fração de indivíduos capacitados no mundo que não têm acesso a uma educação de qualidade ainda é substancial, mas vem diminuindo.

Avanços na farmacologia podem resultar em algum ganho cognitivo ao longo das próximas décadas. Mas depois que os ajustes mais fáceis fossem alcançados — talvez aumentos sustentáveis em energia mental e habilidade de concentração, e

um controle melhor da taxa de consolidação da memória de longo prazo —, ganhos subsequentes serão cada vez mais difíceis de obter. Diferentemente das abordagens da alimentação e da saúde pública, melhorar a cognição por meio de drogas inteligentes pode ser mais fácil antes de se tornar mais difícil. O campo da neurofarmacologia ainda não possui muitos dos conhecimentos básicos necessários para efetuar uma intervenção competente num cérebro saudável. Isso se deve parcialmente ao fato de que a medicina dedicada ao aprimoramento das capacidades cognitivas tem sido negligenciada como uma área legítima de pesquisa. Se a neurociência e a farmacologia continuarem a progredir por mais tempo sem se concentrar no aprimoramento cognitivo, talvez, então, alguns ganhos relativamente fáceis pudessem ser obtidos quando finalmente o desenvolvimento dos nootrópicos se tornasse uma prioridade séria.[195]

O aprimoramento cognitivo genético tem um perfil de resistência em formato de U, similar ao dos nootrópicos, porém com um potencial de ganho maior. A resistência se mostrará alta enquanto o único método disponível for o de cruzamento seletivo mantido por muitas gerações, algo que é obviamente difícil de alcançar em uma escala globalmente significativa. O aprimoramento genético se tornará mais fácil na medida em que for desenvolvida uma tecnologia para testes e seleções genéticas efetivas com baixo custo (e particularmente quando o processo iterativo de seleção de embriões se tornar factível para seres humanos). Essas novas técnicas possibilitarão a exploração do conjunto de variações genéticas humanas existentes com o objetivo de encontrar alelos que aprimorem a inteligência. Entretanto, conforme os melhores alelos existentes sejam incorporados aos pacotes de aprimoramento genético, ganhos posteriores serão difíceis de obter. A necessidade de abordagens mais inovadoras para a modificação genética poderá aumentar a resistência. Há limites para a rapidez do progresso através do caminho do aprimoramento genético, mais notadamente o fato de que as intervenções nas linhas germinais estão sujeitas a um inevitável atraso de maturação: isso contraria fortemente a possibilidade de uma partida rápida ou moderada.[196] O fato de que a seleção de embriões só pode ser aplicada no contexto de uma fertilização *in vitro* diminuirá a sua taxa de adoção: outro fator limitante.

A resistência em relação à via cérebro-computador parece ser inicialmente muito alta. Num improvável cenário em que, de alguma maneira, se torne fácil inserir implantes cerebrais e alcançar integração funcional de alto nível com o córtex, essa resistência despencaria. A longo prazo, a dificuldade de progredir por essa via seria similar àquela presente no aperfeiçoamento de emulações ou IAs, uma vez que o grosso do sistema de inteligência cérebro-computador estaria situado na parte do computador.

A resistência para tornar redes e organizações mais eficientes de um *modo geral* é alta. Uma quantidade grande de esforço será direcionada para superar essa resistência, e o resultado seria uma melhoria anual da capacidade total da humanidade de apenas alguns pontos percentuais.[197] Além disso, mudanças no ambiente interno e externo fazem com que as organizações, mesmo que tenham sido eficientes em algum momento, logo se tornem inadequadas para as novas circunstâncias. Esforços contínuos de reforma são consequentemente necessários

ao menos para prevenir uma deterioração. Uma mudança brusca na taxa de ganho de eficiência organizacional média talvez seja concebível, mas é difícil antever de que maneira um cenário desse tipo, mesmo o mais radical, seria capaz de produzir resultados mais rapidamente do que um cenário de partida lenta, uma vez que organizações operadas por humanos estão confinadas a trabalhar em uma escala de tempo humana. A internet continua sendo uma fronteira estimulante com muitas oportunidades para o aumento da inteligência coletiva e que apresenta uma resistência, no momento, aparentemente moderada — o progresso é, de alguma forma, rápido, mas muitos esforços são necessários para que esse progresso aconteça. Podemos esperar que esse progresso aumente conforme os ganhos mais simples (tais como ferramentas de busca e e-mail) tenham sido esgotados.

Caminhos de emulação e IA

É difícil estimar a dificuldade do avanço em direção a uma emulação completa do cérebro. Ainda assim, podemos destacar um marco específico no futuro: a emulação bem-sucedida do cérebro de um inseto. Esse marco se situa no alto de uma montanha e a sua conquista revelaria uma visão do terreno adiante, permitindo-nos uma estimativa razoável da resistência em relação ao aumento da tecnologia até que seja possível uma emulação completa do cérebro humano. (Uma emulação bem-sucedida do cérebro de um mamífero pequeno, como um rato, nos colocaria numa posição estratégica que permitiria estimar, com certo grau de precisão, a distância até uma emulação total do cérebro humano.) Ao contrário, o caminho em direção à inteligência artificial pode não apresentar marcos óbvios ou pontos de observação prévia. É totalmente possível que a busca pela inteligência artificial se mantenha aparentemente perdida em uma densa floresta até que uma descoberta inesperada revele a linha final em uma clareira a apenas alguns passos de distância.

Relembre a distinção entre estas duas questões: Quão difícil é alcançar níveis aproximadamente humanos de habilidade cognitiva? E quão difícil seria, partindo disso, alcançar níveis super-humanos? A primeira questão é relevante, principalmente para prever quanto tempo teremos até o início da partida. No entanto, a segunda questão é crucial na avaliação do formato da partida, que é nosso objetivo aqui. Embora possa ser tentador supor que o passo necessário para sair do nível humano e atingir o nível do super-humano deva ser o mais difícil — esse passo, afinal de contas, se dá “em uma altitude maior”, na qual melhorias devem ser adicionadas a um sistema já bastante desenvolvido —, essa seria uma suposição bastante insegura. É bem possível que a resistência *caia* no momento em que a máquina alcançar a paridade com os humanos.

Considere, primeiramente, a emulação completa do cérebro. As dificuldades envolvidas na criação da primeira emulação humana seriam de um tipo bem diferente daquelas relacionadas ao aperfeiçoamento de uma emulação já existente. Criar uma primeira emulação envolve enormes desafios tecnológicos, particularmente em relação ao desenvolvimento das capacidades necessárias para digitalizar e interpretar imagens. Esse passo também pode exigir uma quantidade considerável de capital físico — um parque de máquinas de escala industrial com centenas de escâneres de alta resolução não é algo impossível. Em contrapartida, o

aumento da qualidade de uma emulação já existente envolve o ajuste dos algoritmos e estruturas de dados: essencialmente um problema de software que poderia se mostrar muito mais fácil do que o aperfeiçoamento da tecnologia de imagem necessária para criar um modelo original. Programadores poderiam facilmente experimentar algumas mudanças simples, tais como o aumento da contagem dos neurônios em áreas corticais diferentes para verificar como isso afetaria o desempenho.[198] Eles também poderiam trabalhar na otimização do código e na busca por modelos computacionais mais simples que preservassem a funcionalidade essencial de neurônios individuais ou pequenas redes de neurônios. Se o último pré-requisito tecnológico a surgir for a digitalização ou a tradução, com poder computacional relativamente abundante, então pouca atenção teria sido dada às implementações mais eficientes durante a fase de desenvolvimento, e oportunidades fáceis para ganhos de eficiência computacional estariam disponíveis. (Uma reorganização arquitetural mais fundamental também poderia ser possível, mas isso nos tiraria do caminho da emulação e nos levaria para o território da IA.)

Após a produção da primeira emulação, outra maneira de melhorar a base de código é a digitalização de cérebros adicionais com habilidades e talentos diferentes ou superiores. O crescimento da produtividade também ocorreria como uma consequência da adaptação de estruturas organizacionais e do fluxo de trabalho aos atributos únicos das mentes digitais. Como na economia humana não há precedentes de um trabalhador que possa ser literalmente copiado, reiniciado, operado em diferentes velocidades e assim por diante, os responsáveis pela primeira emulação teriam espaço suficiente para inovações em práticas gerenciais.

Quando a emulação completa do cérebro se tornar possível, a resistência pode voltar a subir após uma queda inicial. Mais cedo ou mais tarde, as ineficiências de implementação mais gritantes terão sido otimizadas, as variações algorítmicas mais promissoras terão sido testadas e as oportunidades mais fáceis para a inovação organizacional terão sido exploradas. O catálogo de modelos terá expandido tanto que a aquisição de uma maior quantidade de imagens digitalizadas dos cérebros poderia trazer poucos benefícios em relação ao trabalho com os modelos já existentes. Como é possível multiplicar um modelo, cada cópia pode ser individualmente treinada para trabalhar em diferentes áreas, o que pode ser feito em uma velocidade eletrônica, e provavelmente a quantidade de cérebros que precisariam ser digitalizados para capturar a maior parte do potencial de ganho econômico seja pequena. Possivelmente um único cérebro seria suficiente.

Outra causa provável do aumento da resistência é a possibilidade de que as emulações ou seus apoiadores biológicos se organizem para apoiar regulamentações que restrinjam o uso de trabalhadores emulados, limitem a cópia de emulações e proíbam certos tipos de experimentos com mentes digitais, instituindo direitos aos trabalhadores, um salário mínimo para emulados e assim por diante. É igualmente possível, todavia, que desenvolvimentos políticos caminhassem na direção oposta, contribuindo para uma queda na resistência. Isso poderia ocorrer se as restrições iniciais ao uso da mão de obra de emulados dessem lugar a uma exploração desenfreada conforme a competição aumente e os

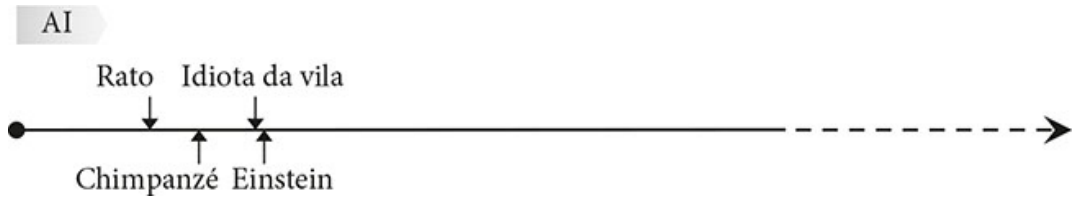
custos econômicos e estratégicos para se ocupar um lugar moralmente mais alto se tornem claros.

Para a inteligência artificial (inteligência de máquina não emulada), a dificuldade de se elevar um sistema do nível humano para a inteligência super-humana por meio de melhorias algorítmicas depende dos atributos do sistema em particular. Diferentes arquiteturas talvez apresentem resistências bem diferentes.

Em algumas situações a resistência poderia ser extremamente baixa. Por exemplo, se a IA de nível humano ainda não existir em virtude da falta de um determinado insight por parte dos programadores, então, quando a descoberta final ocorrer, a IA poderá saltar para um nível radicalmente superior ao humano sem ao menos passar pelos degraus intermediários. Outra situação na qual a resistência seria extremamente baixa é aquela em que um sistema de IA seja capaz de se tornar inteligente por meio de dois modos diferentes de processamento. Para ilustrar essa possibilidade, suponhamos que uma IA é composta de dois subsistemas, um que possui técnicas de solução de problemas de domínio específico e o outro que apresenta habilidades gerais de raciocínio. Esse poderia ser, então, um caso no qual o segundo subsistema se mantém abaixo de um certo nível de capacidade, não contribuindo em nada com o desempenho total do sistema, pois as soluções geradas por ele são sempre inferiores àquelas criadas pelo subsistema de domínio específico. Agora suponha que uma pequena quantidade de poder de otimização seja aplicada ao subsistema de habilidade geral e que isso produza um rápido crescimento na capacidade do subsistema. Primeiramente não observaríamos nenhum aumento no desempenho geral, o que indicaria uma resistência alta. Então, uma vez que o subsistema de capacidade geral ultrapassasse o patamar no qual as soluções geradas por ele superassem aquelas criadas pelo subsistema de domínio específico, o desempenho geral do sistema passaria a apresentar melhoras com a mesma rapidez que o subsistema de propósito geral, mesmo que a quantidade de poder de otimização aplicada permanecesse constante: a resistência do sistema cairia abruptamente.

Também é possível que a nossa tendência natural de encarar a inteligência a partir de uma perspectiva antropocêntrica nos conduza a subestimar as melhorias em sistemas sub-humanos e, assim, superestimar a resistência. Eliezer Yudkowsky, um teórico em IA que escreveu extensivamente sobre o futuro da inteligência de máquina, apresenta a questão da seguinte maneira (ver figura 8):

FIGURA 8



UMA ESCALA MENOS ANTROPOMÓRFICA? O intervalo entre uma pessoa estúpida e uma inteligente pode parecer grande em uma perspectiva antropocêntrica, contudo, em uma abordagem menos paroquial, ambas possuem mentes praticamente indistinguíveis.* Provavelmente, a construção de uma inteligência de máquina com um nível geral de inteligência comparável a um “idiota da vila” será bem mais difícil e levará muito mais

tempo do que o aprimoramento de tal sistema até que ele se torne muito mais inteligente do que qualquer humano.
**A ilustração demonstra a habilidade cognitiva como um parâmetro unidimensional com o objetivo de manter o desenho simples. Mas isso não é essencial para a questão aqui apontada. Seria possível, por exemplo, representar um perfil de habilidade cognitiva com uma hipersuperfície em um espaço multidimensional.*

> *Aparentemente* a IA poderá dar um grande salto em inteligência, simplesmente como resultado do antropomorfismo, a tendência humana de pensar "o idiota da vila" e Einstein como os pontos extremos da escala de inteligência, em vez de percebê-los como pontos quase indistinguíveis em uma escala geral de mentes. Tudo o que for mais estúpido do que um humano estúpido pode parecer para nós um simples "estúpido". Imagina-se a "seta da IA" em uma escala de inteligência crescente, passando pelos ratos e chimpanzés, com as IAs permanecendo "estúpidas" em virtude da incapacidade de falar fluentemente uma língua ou escrever artigos científicos e, então, a IA ultrapassa essa pequena lacuna entre um completo idiota e um super-Einstein no período de um mês ou um período igualmente curto.[199]

A partir dessas diversas considerações, é possível concluir o quão difícil seria prever a complexidade relacionada à criação de melhorias algorítmicas na primeira IA que alcance um nível aproximadamente humano de inteligência geral. Há, pelo menos, algumas circunstâncias possíveis nas quais a resistência algorítmica é baixa. Mas mesmo que a resistência algorítmica seja muito alta, não poderia ser excluída a possibilidade de que a resistência geral dessa IA seja baixa. Pois, além do aperfeiçoamento de seus algoritmos, poderia ser fácil aumentar a inteligência do sistema de outras maneiras. Há outros dois fatores que podem ser aperfeiçoados: conteúdo e hardware.

Primeiro, considere aperfeiçoamentos de conteúdo. Por "conteúdo", nos referimos àqueles componentes dos ativos de software de um sistema que não façam parte da sua arquitetura algorítmica central. O conteúdo pode incluir, por exemplo, bancos de dados que armazenem percepções, bibliotecas de técnicas especializadas e inventários de conhecimento declarativo. Em vários tipos de sistema a distinção entre arquiteturas algorítmicas e conteúdo é muito indefinida. Ainda assim, essa distinção pode ser utilizada como uma forma rápida e aproximada de ressaltar uma fonte potencialmente importante de ganhos de capacidade em inteligência de máquina. Uma maneira alternativa para expressar a mesma ideia é considerar que a capacidade de resolver problemas intelectuais de um sistema pode ser aumentada não apenas por meio do aprimoramento de suas habilidades, mas também com a expansão dos conhecimentos já adquiridos pelo sistema.

Considere um sistema de IA contemporâneo, tal como o TextRunner (um projeto de pesquisa da Universidade de Washington) ou o Watson, da IBM (um sistema que venceu o jogo de perguntas *Jeopardy!*). Esses sistemas podem extrair alguns tipos de informação semântica a partir da análise de texto. Embora eles não sejam capaz de compreender a leitura da mesma forma ou na mesma medida que os humanos, eles podem, entretanto, extrair quantidades significativas de informação da linguagem natural e usá-la para fazer deduções simples e responder perguntas. Eles também podem aprender por experiência, criando representações mais

abrangentes de um conceito conforme encontrem exemplos adicionais de uso. Esses sistemas são projetados para operar sem supervisão durante a maior parte do tempo (ou seja, aprendem estruturas ocultas em dados não rotulados sem indicações de erro ou sinais de recompensa, dispensando orientação humana), além de ser rápidos e escaláveis. O TextRunner, por exemplo, trabalha com um volume de 500 milhões de páginas da web.[200]

Agora imagine um descendente remoto de um sistema desse tipo que tenha adquirido a habilidade de ler com o mesmo nível de compreensão de um humano de dez anos, mas com uma velocidade de leitura similar ao TextRunner. (O que seria, provavelmente, um problema IA-completo.) Então, estamos imaginando um sistema que pensa muito mais rápido e que possui uma memória muito superior a de um humano adulto, embora saiba muito menos, e talvez o efeito concreto desse cenário seja um sistema aproximadamente equivalente a um humano na sua capacidade geral de resolução de problemas. Mas sua resistência de conteúdo é muito baixa — baixa o suficiente para provocar uma partida. Em algumas semanas, o sistema teria lido e dominado todo o conteúdo da Biblioteca do Congresso. Então o sistema saberia muito mais do que qualquer ser humano e pensaria muitíssimo mais rápido: ele teria se tornado (ao menos) fracamente superinteligente.

Um sistema pode, dessa maneira, aumentar consideravelmente sua capacidade intelectual efetiva por meio da absorção de conteúdos pré-produzidos acumulados durante séculos de ciência e civilização humana: por exemplo, através da leitura da internet. Se uma IA alcança o nível humano sem ter tido previamente acesso a esse material ou sem ter sido capaz de processá-lo, então a resistência geral da IA será baixa, mesmo que seja difícil aperfeiçoar suas arquiteturas algorítmicas.

A resistência de conteúdo também é um conceito relevante para as emulações. Uma emulação de alta velocidade é vantajosa não apenas por ser capaz de completar as mesmas tarefas mais depressa do que um humano biológico, mas também porque pode acumular conteúdo rapidamente quando necessário, como o conhecimento e as habilidades relevantes para uma tarefa específica. Entretanto, para alcançar o potencial total de uma acumulação rápida de conteúdo, um sistema precisa ter uma capacidade proporcionalmente grande de memória. Há pouco sentido em ler um dicionário inteiro se você já tiver esquecido tudo sobre o significado de “aba” assim que chegar em “adaga”. Embora um sistema de IA muito provavelmente possua capacidade de memória adequada, as emulações herdariam algumas das limitações de capacidade dos modelos humanos. Elas consequentemente necessitariam de melhorias na sua arquitetura para se tornarem aptas a uma aprendizagem ilimitada.

Até agora temos considerado a resistência da arquitetura e do conteúdo, ou seja, o quão difícil seria aprimorar o *software* de uma inteligência de máquina que tenha alcançado a paridade com os humanos. Olhemos agora para uma terceira maneira de aperfeiçoar o desempenho da inteligência da máquina: aprimorar seu hardware. Qual seria a resistência para melhorias de hardware?

Iniciando com um software inteligente (uma emulação ou uma IA), é possível ampliar a *inteligência coletiva* simplesmente por meio do uso de computadores adicionais para rodar mais instâncias do programa.[201] Outra forma de ampliar a

inteligência rápida seria transferir o programa para computadores mais velozes. Dependendo do grau de paralelização permitido pelo programa, a inteligência rápida poderia ser amplificada através da execução do programa em mais processadores. É mais provável que isso seja factível para emulações cuja arquitetura seja altamente paralelizada, mas vários programas de IA também possuem sub-rotinas importantes que podem se beneficiar de uma paralelização massiva. A amplificação da *inteligência de qualidade* através do aumento de poder computacional também pode ser possível, mas esse é um caso menos direto.[202]

É bem provável que a resistência para ampliar a inteligência coletiva ou a inteligência rápida (e possivelmente a inteligência de qualidade) em um sistema com software de nível humano seja baixa. A única dificuldade seria obter um maior poder computacional. Há várias maneiras para que um sistema possa expandir sua base de hardware, sendo cada uma delas relevante de acordo com uma escala de tempo diferente.

No curto prazo, o poder computacional deve aumentar mais ou menos proporcionalmente aos recursos financeiros disponíveis para o projeto: o dobro de recursos compraria o dobro de computadores, possibilitando que o dobro de instâncias do software rode simultaneamente. O surgimento dos serviços de computação em nuvem oferece ao projeto a opção de aumentar seus recursos computacionais sem nem ao menos ter que esperar que os novos computadores sejam entregues e instalados, embora preocupações com o sigilo possam favorecer o uso de computadores internos. (Em determinados cenários, o poder computacional poderia ser obtido de outras formas, como através do comando de *botnets*.)[203] O quão fácil seria aumentar a capacidade de um sistema depende da quantidade de poder computacional que o sistema inicialmente utiliza. Um sistema que inicialmente roda em um PC poderia ter sua capacidade computacional aumentada milhares de vezes com míseros 1 milhão de dólares. Já um programa que rode em um supercomputador geraria muito mais custos para ter sua capacidade aumentada.

Considerando-se um período de tempo ligeiramente mais longo, o custo de aquisição de hardware adicional pode encarecer conforme uma porção crescente da capacidade computacional global instalada seja usada para rodar mentes digitais. Por exemplo, em um cenário de emulação baseado em um mercado competitivo, o custo para rodar uma cópia adicional de uma emulação deve subir até ficar ligeiramente igual ao rendimento gerado pela cópia marginal, enquanto investidores cobrem o valor da infraestrutura de computação existente para se equiparar com o retorno que eles esperam obter com seus investimentos (porém se apenas um projeto tiver dominado a tecnologia, pode ser que ele ganhe um grau de força de monopsônio no mercado de poder computacional e, então, o custo seria mais baixo).

Numa escala de tempo um pouco mais longa, a oferta de poder computacional crescerá conforme a nova capacidade seja instalada. Um pico de demanda impulsionaria a produção em instalações existentes para a fabricação de semicondutores e estimularia a construção de novas fábricas. (Um aumento de performance pontual, talvez com uma a duas ordens de magnitude, pode também

ser obtido através da utilização de microprocessadores customizados.)[204] Sobretudo, a onda crescente de melhorias na tecnologia fornecerá volumes cada vez maiores de poder computacional para ser consumido pelas turbinas das máquinas pensantes. Historicamente, a taxa de progresso em tecnologia computacional tem sido descrita pela famosa lei de Moore, a qual, em uma de suas variações, afirma que a capacidade computacional por dólar investido duplica mais ou menos a cada dezoito meses.[205] Embora seja impossível garantir que essa taxa de progresso seja mantida até o desenvolvimento de uma inteligência de máquina de nível humano, enquanto os limites físicos fundamentais ainda não forem alcançados, ainda haverá espaço para avanços na tecnologia computacional.

Há, então, razões para supor que a resistência de hardware não será muito alta. A compra de mais poder computacional para o sistema, uma vez que ele prove seu valor alcançando a inteligência de nível humano, pode facilmente adicionar várias ordens de magnitude de poder computacional (dependendo de quão modesto era o hardware utilizado pelo projeto antes da expansão). A customização de processadores poderia adicionar uma ou duas ordens de magnitude. Outros meios de expansão da base de hardware, tais como a construção de mais fábricas e avanços nos limites da tecnologia computacional, levam mais tempo — normalmente vários anos, embora esse intervalo possa ser radicalmente compactado se a superinteligência de máquina for capaz de revolucionar a produção e o desenvolvimento de tecnologia.

Em resumo, podemos falar sobre a probabilidade de um *excedente de hardware*: quando da criação de um software de nível humano, um poder computacional suficiente pode estar disponível para rodar um número grande de cópias em uma grande velocidade. A resistência de software, como foi discutido acima, é difícil de ser avaliada, mas pode ser inclusive mais baixa do que a resistência de hardware. Em particular, pode haver um *excedente de conteúdo* na forma de conteúdo pré-existente (por exemplo, a internet) que se tornará disponível para um sistema quando este alcançar a paridade com os humanos. Um *excedente algorítmico* — melhorias algorítmicas desenvolvidas previamente — também é possível, mas talvez menos provável. Melhorias de software (tanto algorítmicas quanto de conteúdo) podem oferecer ganhos potenciais de desempenho de algumas ordens de magnitude, os quais poderiam ser relativamente fáceis de acessar quando uma mente digital alcançasse paridade com os humanos, além de ganhos de desempenho que poderiam ser obtidos com o uso de mais hardware ou com um hardware melhor.

Poder de otimização e explosividade

Após examinar a questão de resistência, devemos nos voltar agora para a outra metade da nossa equação esquemática: *poder de otimização.* Relembrando: *taxa de mudança da inteligência = poder de otimização/resistência.* Conforme refletido nessa equação, uma partida rápida não requer que a resistência durante a fase de transição seja baixa. Uma partida rápida também poderia acontecer caso a resistência fosse constante ou até mesmo se estivesse aumentando

moderadamente, contanto que o poder de otimização aplicado no aperfeiçoamento do sistema aumente suficientemente rápido. Como veremos agora, há bom embasamento para considerar que o poder de otimização aplicado *aumentará* durante a transição, pelo menos na ausência de medidas deliberadas para evitar que isso aconteça.

Podemos distinguir duas fases. A primeira começa com o início da partida, quando o sistema alcança a linha de referência humana para a inteligência individual. Conforme as capacidades do sistema continuem a aumentar, ele pode se utilizar de alguma ou de todas essas capacidades para aprimorar a si mesmo (ou desenvolver um novo sistema — o que, para nossa argumentação, significa a mesma coisa). Todavia, a maior parte do poder de otimização aplicado ao sistema ainda partirá de fora do sistema, do trabalho dos programadores e engenheiros envolvidos no projeto ou de trabalhos realizados em outras partes do mundo que possam ser apropriados ou utilizados no projeto.[206] Se essa fase se arrastar por qualquer período de tempo significativo, podemos esperar que a quantidade de poder de otimização aplicado ao sistema crescerá. Influências tanto internas como externas provavelmente aumentariam na medida em que a promessa de sucesso da abordagem escolhida se tornasse evidente. Pesquisadores poderiam trabalhar mais arduamente, mais pesquisadores poderiam ser contratados e poder computacional adicional poderia ser adquirido para acelerar o progresso. O aumento poderia ser especialmente dramático se o desenvolvimento de uma inteligência de máquina de nível humano tomasse o mundo de surpresa. Nesse caso o que era previamente um pequeno projeto de pesquisa poderia então, subitamente, se tornar o foco de intensa pesquisa e desenvolvimento ao redor do mundo (embora alguns desses esforços possam estar ligados a diferentes projetos que estejam competindo entre si).

Uma segunda fase de crescimento começará se, em algum momento, o sistema tiver adquirido a capacidade necessária para que a maior parte do poder de otimização exercido surja do próprio sistema (marcado pelo nível da variável “cruzamento” na figura 7). Isso muda fundamentalmente a dinâmica, pois qualquer aumento na capacidade do sistema agora se traduzirá em um aumento proporcional da quantidade de poder de otimização que estivesse sendo aplicado para melhorias posteriores. Se a resistência continuar constante, essa realimentação dinâmica produzirá um crescimento exponencial (ver quadro 4). O tempo de duplicação depende do cenário, mas pode ser extremamente curto — meros segundos em alguns cenários — se o crescimento estiver acontecendo em velocidades eletrônicas, o que pode ser resultado de aperfeiçoamentos algorítmicos ou do aproveitamento de um excedente de conteúdo ou de hardware.[207] O crescimento movido por construções físicas, tais como a produção de novos computadores ou equipamentos de produção, exigiria uma escala de tempo um pouco mais longa (que ainda assim poderia ser muito curta se comparada com a taxa de crescimento atual da economia mundial).

Portanto, é provável que o poder de otimização aplicado aumentará durante a transição: inicialmente porque humanos realmente tentam melhorar uma inteligência de máquina que se mostre bastante promissora, e posteriormente porque a própria

inteligência de máquina se torna capaz de promover progressos futuros em velocidades digitais. Isso criaria a possibilidade real de uma partida rápida ou moderada, *inclusive se a resistência fosse constante ou aumentasse ligeiramente perto da linha de referência humana.*[208] Entretanto, vimos na subseção anterior que existem fatores que poderiam conduzir a uma grande queda na resistência perto do nível de capacidade da linha de referência humana. Esses fatores incluem, por exemplo, a possibilidade de expansão rápida do hardware, uma vez que uma mente de software funcional tenha sido criada; a possibilidade de melhorias algorítmicas; a possibilidade de digitalização cerebral adicional (no caso de uma emulação completa do cérebro) e a possibilidade de incorporar rapidamente uma ampla quantidade de conteúdo através da leitura da internet (no caso de uma inteligência artificial).[209]

Independentemente dessas observações, o formato da curva de resistência na região relevante ainda não está bem caracterizado. Em particular, não está claro o quão difícil seria melhorar a qualidade de software de uma emulação de nível humano ou de uma IA. A dificuldade de expansão do poder de hardware disponível para o sistema também não está clara. Enquanto hoje seria relativamente fácil aumentar o poder computacional disponível para um pequeno projeto gastando mil vezes mais em poder computacional ou esperando alguns anos para que o preço dos computadores caia, é possível que a primeira inteligência de máquina a alcançar o nível humano seja resultado de um grande projeto que envolvesse supercomputadores caros, os quais não poderiam ser replicados de forma barata, e que a lei de Moore já tenha deixado de valer. Por essas razões, apesar de uma partida moderada ou rápida parecer mais provável, a possibilidade de uma partida lenta não pode ser excluída.[210]

QUADRO 4

Sobre a cinética de uma explosão de inteligência

A taxa de mudança da inteligência é dada pela razão entre o poder de otimização aplicado ao sistema e a resistência do sistema:

$$\frac{dI}{dt} = \frac{\mathfrak{O}}{\mathfrak{R}}.$$

A quantidade de poder de otimização atuando em um sistema é a soma de qualquer poder de otimização que o sistema exerce nele mesmo e do poder de otimização exercido de fora. Por exemplo, uma IA embrionária pode ser melhorada por meio de uma combinação de seus próprios esforços e dos esforços de uma equipe de programadores humanos, e talvez também dos esforços de uma comunidade mais ampla de pesquisadores que trabalhem gerando avanços contínuos na indústria de semicondutores, ciência da computação e campos afins:[211]

$$O = O_{sitema} + O_{projeto} + O_{mundo}.$$

Uma IA embrionária inicia com capacidades cognitivas muito limitadas. No início, portanto, $O_{sistema}$ é pequeno.[212] E com relação a $O_{projeto}$ e O_{mundo}? Há casos nos quais um único projeto tem mais capacidade relevante do que o resto do mundo todo em conjunto — o Projeto Manhattan, por exemplo, convocou grande

parte dos melhores físicos do mundo para Los Alamos a fim de que trabalhassem na bomba atômica. Mais comumente, qualquer projeto em particular conta somente com uma pequena fração do total da capacidade relevante de pesquisa no mundo. Mesmo quando o mundo exterior tem um total maior de capacidade relevante de pesquisa do que qualquer projeto em particular, $O_{projeto}$ ainda assim pode exceder O_{mundo}, desde que a maior parte da capacidade do mundo exterior não esteja focada no sistema particular em questão. Se um projeto começa a parecer promissor — o que acontecerá quando um sistema ultrapassar a linha de referência humana, ou até antes —, ele pode atrair investimentos adicionais, aumentando $O_{projeto}$. Se as conquistas do projeto forem públicas, O_{mundo} também pode aumentar, conforme o progresso inspire maior interesse em inteligência de máquina em geral e várias forças se esforcem para entrar no jogo. Durante a fase de transição, então, o poder de otimização total aplicado para melhorar um sistema cognitivo provavelmente aumentará conforme o crescimento da capacidade do sistema.[213]

Conforme as capacidades do sistema cresçam, é possível chegar a um ponto no qual o poder de otimização gerado pelo próprio sistema comece a ser maior do que o poder de otimização nele aplicado por forças exteriores (para todas as dimensões significativas de melhoria):

$$O_{sitema} > O_{projeto} + O_{mundo}.$$

Esse cruzamento é significativo porque, além desse ponto, melhorias adicionais das capacidades do sistema contribuem fortemente para o aumento do poder total de otimização aplicado na melhoria do sistema. Desse modo, entramos em um regime de forte automelhoria recursiva. Isso conduzirá a um crescimento explosivo da capacidade do sistema sob uma gama consideravelmente ampla de diferentes formatos da curva de resistência.

Para ilustrar, considere primeiro um cenário no qual a resistência é constante, de forma que a taxa de aumento da inteligência de uma IA seja igual ao poder de otimização que está sendo aplicado. Considere que todo poder de otimização aplicado venha da própria IA e que a IA aplica toda a sua inteligência na tarefa de amplificar sua própria inteligência, de tal forma que $O_{sistema} = I$.[214] Então teremos

$$\frac{dI}{dt} = \frac{I}{k}.$$

A solução dessa equação diferencial simples gera a função exponencial

$$I = Ae^{t/k}$$

Porém a resistência constante é um caso muito especial. A resistência pode muito bem baixar perto da linha de referência humana, em virtude de um ou mais dos fatores mencionados na subseção anterior, continuando baixa por volta do cruzamento e um pouco mais além (talvez até que o sistema se aproxime dos limites físicos fundamentais). Por exemplo, suponhamos que o poder de otimização aplicado ao sistema seja razoavelmente constante (ou seja, $O_{projeto} + O_{mundo} \approx c$) antes que o sistema se torne capaz de contribuir substancialmente para o seu próprio design, e que isso faça com que o sistema duplique sua capacidade a cada dezoito meses. (Isso estaria aproximadamente de acordo com a taxa de progresso da lei de Moore associada aos avanços de software.)[215] Essa taxa de progresso, se alcançada através de um poder de otimização mais ou menos constante, implica na diminuição da resistência de forma inversamente proporcional ao poder do sistema:

$$\frac{dI}{dt} = \frac{c}{1/I} = cI.$$

Se a resistência continuar a cair de acordo com esse padrão hiperbólico, quando a IA alcançar o ponto de cruzamento a quantidade total de poder de otimização aplicado para melhorar a IA terá dobrado. Então, teremos:

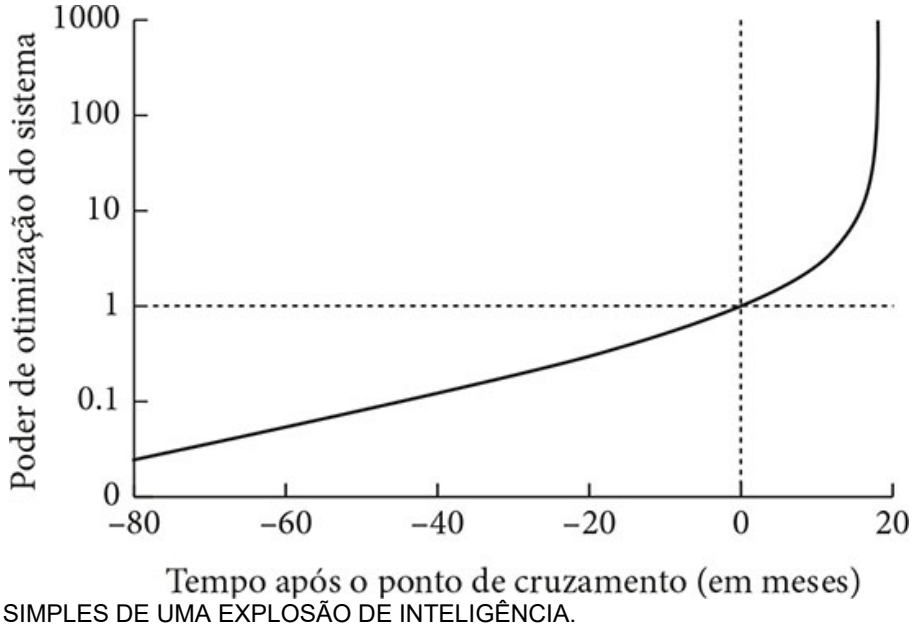
$$\frac{dI}{dt} = \frac{(c+I)}{1/I} = (c+I)I.$$

A próxima duplicação ocorrerá 7,5 meses mais tarde. Dentro de 17,9 meses, a capacidade do sistema terá crescido milhares de vezes, atingindo, assim, uma superinteligência rápida (figura 9).

Essa trajetória particular de crescimento tem uma singularidade positiva em t = 18 meses. Na realidade, a hipótese de que a resistência é constante deixaria de existir conforme o sistema começasse a se aproximar dos limites físicos do processamento de informação, senão antes.

Esses dois cenários são apenas ilustrativos; muitas outras trajetórias são possíveis, dependendo do formato da curva da resistência. Podemos simplesmente afirmar que a forte realimentação estabelecida por volta do ponto de cruzamento tende a fazer com que a partida seja mais rápida do que seria em outra situação.

FIGURA 9



UM MODELO SIMPLES DE UMA EXPLOSÃO DE INTELIGÊNCIA.

05

CAPÍTULO 5

VANTAGEM ESTRATÉGICA DECISIVA

Uma questão distinta, embora relacionada à questão da cinética, é se haverá apenas uma ou muitas forças superinteligentes. Uma explosão de inteligência poderia lançar um projeto tão à frente de todos os outros a ponto de torná-lo capaz de ditar o futuro? Ou o progresso será mais uniforme, desenrolando-se em uma frente ampla, com muitos projetos em andamento sem que nenhum deles garanta uma liderança expressiva e permanente?

O capítulo anterior analisou um parâmetro crucial na determinação da distância que poderia plausivelmente ser criada entre uma força de liderança e seus concorrentes mais próximos — especificamente, a velocidade de transição da inteligência humana para uma inteligência fortemente super-humana. Isso sugere uma análise preliminar. Se a partida for *rápida* (realizada ao longo de horas, dias ou semanas), então é pouco provável que dois projetos independentes existam concomitantemente: quase certamente, o primeiro projeto teria completado a sua partida antes que qualquer outro projeto tivesse iniciado a sua. Se a partida for *lenta* (estendendo-se por muitos anos ou décadas), então é plausível que existam múltiplos projetos realizando partidas concomitantemente, de modo que, embora os projetos ao final da transição pudessem ter aumentado em muito suas capacidades, em nenhum momento qualquer um dos projetos teria estado tão à frente a ponto de alcançar uma liderança expressiva. Uma partida de velocidade *moderada* estaria sujeita a ambas as possibilidades: poderia ou não haver mais de um projeto realizando sua partida simultaneamente.[216]

Um projeto de máquina inteligente poderia se destacar tanto em relação aos seus competidores a ponto de conseguir uma *vantagem estratégica decisiva* — ou seja, um nível de tecnologia e outras vantagens suficientes para que ele consiga um domínio completo do mundo? Se um projeto conseguisse obter uma vantagem estratégica decisiva, ele a usaria para bloquear os concorrentes e formar um *singleton* (uma ordem mundial onde haveria uma única agência para tomar decisões a nível global)? E caso houvesse um projeto vencedor, quão “grande” ele seria — não em termos de tamanho físico ou orçamento, mas em termos da quantidade de pessoas responsáveis pelo controle da sua criação? Consideraremos essas questões a seguir.

O projeto líder terá uma vantagem estratégica decisiva?

Numa corrida, um fator que influencia a distância entre quem corre na frente e os que vêm atrás é a taxa de difusão daquilo que fornece ao líder uma vantagem competitiva. O corredor da frente talvez possa achar difícil conquistar e manter uma larga vantagem se seus seguidores puderem facilmente copiar suas ideias e inovações. As imitações criam uma corrente de vento contrária que tira a vantagem do líder e beneficia os retardatários, especialmente se a propriedade intelectual não estiver muito bem protegida. O líder também pode ficar especialmente vulnerável à expropriação, à tributação ou poderá ser desmembrado em virtude de alguma regulamentação antimonopólio.

Seria um erro, no entanto, supor que essa corrente de vento contrária aumente monotonicamente de acordo com o espaço entre o líder e seus seguidores. Assim como um ciclista que fica muito para trás dos demais competidores já não é mais protegido do vento pelos ciclistas da frente, um seguidor de tecnologia suficientemente atrasado em relação à linha de frente pode ter dificuldade de assimilar os avanços mais recentes.[217] A lacuna em conhecimento e capacidade pode ter crescido demais. O líder pode ter migrado para uma plataforma tecnológica mais avançada, fazendo com que inovações subsequentes sejam intransferíveis para as plataformas primitivas utilizadas pelos retardatários. Um líder proeminente o suficiente pode ser hábil o bastante para conter o vazamento de informação dos seus programas de pesquisa e de suas instalações sensíveis, ou para sabotar os esforços de seus concorrentes em desenvolver suas próprias capacidades avançadas.

Se o líder for um sistema de IA, ele poderia ter atributos que facilitassem a expansão de suas capacidades e ao mesmo tempo reduzisse sua taxa de difusão. Em uma organização gerenciada por humanos, economias de escala são neutralizadas por ineficiências burocráticas e conflitos de interesse, incluindo dificuldades para manter segredos de negócios.[218] Esses problemas provavelmente limitariam o crescimento de um projeto de inteligência de máquina enquanto fosse operado por humanos. Um sistema de IA, entretanto, talvez possa evitar algumas dessas deseconomias de escala, uma vez que os módulos de IA (ao contrário dos trabalhadores humanos) não precisam ter preferências individuais que estejam em desacordo com aquelas do sistema como um todo. Dessa maneira, o sistema de IA poderia evitar uma quantidade considerável de ineficiências causadas por conflitos de interesse em empreendimentos humanos. A mesma vantagem — possuir partes perfeitamente leais — também facilitaria para um sistema de IA a busca de objetivos secretos de longo alcance. Uma IA não teria empregados descontentes prontos para ser recrutados pela concorrência ou subornados para se tornar informantes.[219]

Por meio da observação de alguns exemplos históricos, podemos ter uma ideia da distribuição de distâncias plausíveis entre tempos de desenvolvimento (ver quadro 5). Atrasos na faixa de poucos meses a alguns anos parecem ser típicos de projetos tecnológicos estrategicamente significativos.

QUADRO 5

Corridas tecnológicas: alguns exemplos históricos

No decorrer de longos períodos históricos, a intensidade com que o conhecimento e a tecnologia se difundem ao redor do mundo tem aumentado. Como resultado, as distâncias temporais entre os líderes em tecnologia e os seguidores mais próximos vêm diminuindo.

A China conseguiu manter um monopólio da produção de seda por mais de 2 mil anos. Achados arqueológicos sugerem que a produção pode ter se iniciado por volta de 3000 a.C. ou mesmo antes.[220] A sericultura era mantida sob sigilo e a revelação de suas técnicas era punida com a morte. A mesma punição era reservada àquele que levasse o bicho-da-seda ou seus ovos para fora da China. Os romanos, apesar do alto preço da seda importada em seu império, nunca aprenderam a arte da manufatura da seda. Por volta de 300 d.C., uma expedição japonesa conseguiu capturar alguns ovos do bicho-da-seda e quatro garotas chinesas, que foram forçadas a revelar a arte aos seus raptores.[221] Os bizantinos se tornaram parte do grupo de produtores em 522 d.C. A história da fabricação da porcelana também mostra longos atrasos. A arte era praticada na China durante a dinastia Tang, por volta de 600 d.C. (e é possível que já estivesse em uso desde 200 d.C.), mas foi dominada pelos europeus apenas no século XVIII.[222] Veículos com rodas apareceram em muitos lugares na Europa e na Mesopotâmia por volta de 3500 a.C., mas chegaram nas Américas somente no período pós-colombiano.[223] Numa escala ainda maior, a espécie humana levou dezenas de milhares de anos para se espalhar por quase todo o globo terrestre, a Revolução Agrícola levou milhares de anos, a Revolução Industrial precisou de apenas centenas de anos, e podemos dizer que uma Revolução da Informação se espalhou globalmente no curso de décadas — embora, é claro, essas transições não sejam necessariamente de igual profundidade. (O videogame *Dance Dance Revolution* se espalhou do Japão para a Europa e depois para a América do Norte em apenas um ano!)

As competições tecnológicas têm sido extensivamente estudadas, particularmente nos contextos das corridas de patentes e de armas.[224] Está além do escopo de nossas investigações revisar essa literatura aqui. Todavia, é instrutivo observar alguns exemplos de corridas tecnológicas estrategicamente significativas do século XX (ver tabela 7).

A respeito dessas seis tecnologias consideradas como estrategicamente importantes pelas superpotências rivais em virtude do seu valor militar ou simbólico, os espaços entre o líder e os seguidores mais próximos eram (muito aproximadamente) 49 meses, 36 meses, quatro meses, um mês, quatro meses e sessenta meses, respectivamente — mais longos do que a duração de uma partida rápida e mais curtos do que a duração de uma partida lenta.[225] Em muitos casos, os projetos dos seguidores se beneficiaram de espionagem e de informações disponíveis publicamente. A mera demonstração da viabilidade de uma invenção também pode encorajar outros a desenvolvê-la de forma independente, e o medo de ficar para trás pode estimular esforços de se equiparar ao rival.

Talvez mais próximas ao caso de uma IA sejam as invenções matemáticas que não exigem o desenvolvimento de uma nova infraestrutura física. Estas são frequentemente publicadas na literatura acadêmica e podem ser, portanto, consideradas universalmente disponíveis; mas em alguns casos, quando a descoberta parece oferecer uma vantagem estratégica, a publicação pode ser adiada. Por exemplo, duas das ideias mais importantes em criptografia de chave pública são o protocolo de troca de chave de Diffie-Hellman e o esquema de codificação RSA. Elas foram descobertas pela comunidade acadêmica em 1976 e 1978, respectivamente, porém mais tarde se confirmou que eram conhecidas pelos criptógrafos do grupo de segurança de comunicações do Reino Unido desde o início dos anos 1970.[226] Grandes projetos de software podem proporcionar uma analogia mais próxima a projetos de IA, embora seja mais difícil fornecer exemplos claros de atrasos típicos pois softwares geralmente são lançados em versões incrementais e as funcionalidades de sistemas concorrentes não são, em geral, diretamente comparáveis.

TABELA 7

Algumas corridas tecnológicas estrategicamente significativas

	Estados Unidos	União Soviética	Reino Unido	França	China	Índia	Israel	Paquistão	Coreia do Norte	África do Sul
Bomba de fissão	1945	1949	1952	1960	1964	1974	1979?	1998	2006	1979?
	1952	1953[227]	1957	1968	1967	1998	?	—	—	—

Bomba de fusão										
Capacidade de lança mento de satélites	1958	1957	1971	1965	1970	1980	1988	—	1998?[228]	—[229]
Capacidade de lança mento de humanos	1961	1961	—	—	2003	—	—	—	—	—
ICBM[230]	1959	1960	1968[231]	1985	1971	2012	2008	—[232]	2006	—[233]
MIRV[234]	1970	1975	1979	1985	2007	2014[235]	2008?			

É possível que a globalização e o aumento da vigilância sejam capazes de reduzir as distâncias típicas entre projetos tecnológicos concorrentes. Ainda assim é provável que haja um limite inferior para o quão curta a diferença de tempo média poderia se tornar (na falta de uma coordenação deliberada).[236] Mesmo na ausência de dinâmicas que causem efeitos de bola de neve, alguns projetos contarão com uma equipe de pesquisa, liderança e infraestrutura melhores ou apenas podem acabar se deparando com melhores ideias. Se dois projetos desenvolvem abordagens distintas e uma delas se mostra melhor, o projeto rival pode levar muitos meses para mudar sua abordagem, mesmo que fosse capaz de monitorar de perto o que o outro projeto estivesse fazendo.

Combinando essas observações com a nossa discussão anterior sobre a velocidade da partida, podemos concluir que é pouco provável que dois projetos estejam próximos o suficiente para experimentar simultaneamente uma partida rápida; no caso de uma partida moderada, isso poderia não ocorrer; já em uma partida lenta, seria bastante provável que muitos projetos passassem pelo processo paralelamente. Mas a análise deve ir além. A questão central não é saber quantos projetos poderiam passar por uma partida ao mesmo tempo, mas quantos emergiriam dessa transição com capacidades suficientemente semelhantes, de modo que nenhum deles possa obter uma vantagem estratégica decisiva. Se o processo de partida começasse de forma relativamente lenta e se tornasse mais rápido, a distância entre os projetos concorrentes tenderia a crescer. Retornando à nossa metáfora da bicicleta, a situação seria análoga a um par de ciclistas que, separados por uma certa distância, sobe um morro íngreme — o espaço entre os dois então se expandiria no momento em que o ciclista da frente alcançasse o topo do morro e começasse a acelerar colina abaixo.

Considere o seguinte cenário de partida moderada. Suponha que um projeto leve um ano para aumentar sua capacidade de IA da linha de referência humana para uma superinteligência forte, e que um projeto entre nessa fase de partida com uma vantagem de seis meses em relação ao próximo projeto mais avançado. Os dois

projetos estariam realizando uma partida simultaneamente. Pode parecer, então, que nenhum dos projetos possui uma vantagem estratégica decisiva. Mas não precisa ser assim. Suponha que sejam necessários nove meses para que seja possível sair da linha de referência humana e atingir o ponto de cruzamento, e outros três meses para avançar desse ponto em direção a uma superinteligência forte. O projeto líder atinge então a superinteligência forte três meses antes do projeto seguinte, sem que este tenha ao menos alcançado o ponto de cruzamento. Isso daria ao projeto líder uma vantagem estratégica decisiva e a oportunidade de transformar sua liderança em controle permanente por meio da desativação dos projetos concorrentes e o estabelecimento de um *singleton*. (Observe que o conceito de *singleton* é abstrato: um *singleton* poderia ser uma democracia, uma tirania, uma única IA dominante, um conjunto sólido de normas globais que incluam cláusulas efetivas para a sua própria aplicação, ou até um soberano extraterrestre — sendo que o aspecto que o define é simplesmente ser uma agência capaz de solucionar todos os problemas principais de coordenação global. Embora não necessariamente, ele pode se assemelhar a alguma forma familiar de governo humano.)[237]

Como há uma grande possibilidade de explosão de crescimento logo após o ponto de cruzamento, quando a forte realimentação positiva de poder de otimização é iniciada, um cenário desse tipo é uma possibilidade concreta e aumenta as chances de que o projeto líder possa obter uma vantagem estratégica decisiva ainda que a partida não seja rápida.

Qual será o tamanho de um projeto bem-sucedido?

Alguns caminhos para a superinteligência exigem muitos recursos e, por isso, serão provavelmente mantidos por projetos com grande financiamento. A emulação completa do cérebro, por exemplo, requer diferentes tipos de técnicas e uma série de equipamentos. Melhorias na inteligência biológica e interfaces cérebro-computador também dependem fortemente da escala do projeto: enquanto uma pequena firma de biotecnologia poderia inventar uma ou duas drogas, alcançar a superinteligência ao longo de um desses caminhos (se for algo realmente possível) exigiria provavelmente muitas invenções e testes variados, e, portanto, o apoio de um setor industrial ou de um programa nacional bem financiado. Conseguir uma superinteligência coletiva através de organizações e redes mais eficientes demanda ainda mais recursos, o que envolveria grande parte da economia mundial.

O caminho da IA é mais difícil de avaliar. É possível que ele exija um amplo programa de pesquisa, mas talvez possa ser percorrido por um pequeno grupo. O cenário de um hacker solitário também não pode ser excluído. Construir uma IA embrionária pode demandar descobertas e algoritmos desenvolvidos ao longo de muitas décadas pela comunidade científica ao redor do mundo. Mas é possível que a descoberta central venha de um único indivíduo ou de um pequeno grupo que consiga encaixar todas as peças. Esse cenário é menos realístico para algumas arquiteturas de IA do que para outras. É provável que um sistema com um grande número de partes que precisem ser ajustadas para trabalhar em conjunto de

maneira eficiente, além de ser cuidadosamente alimentadas com conteúdo cognitivo feito sob medida, exija um projeto maior. Mas se uma IA embrionária puder ser criada na forma de um sistema simples, cuja construção dependa apenas da aplicação de alguns princípios básicos corretos, esse feito talvez esteja ao alcance de uma equipe pequena ou de um único indivíduo. A probabilidade de que a descoberta final seja feita por um projeto pequeno é aumentada caso a maior parte do progresso prévio na área tenha sido publicada ou esteja disponível na forma de software livre.

Devemos fazer distinções entre a questão do tamanho do projeto que *desenvolve* diretamente o sistema e a questão do tamanho do grupo que *controla* se, como e quando o sistema deve ser criado. A bomba atômica foi criada essencialmente por um grupo de cientistas e engenheiros. (O Projeto Manhattan empregou cerca de 130 mil pessoas no seu auge, cuja grande maioria era composta de trabalhadores da construção civil e técnicos.)[238] Esses especialistas técnicos, entretanto, eram controlados pelas Forças Armadas norte-americanas, as quais eram dirigidas pelo governo norte-americano, que, em última instância, prestava contas ao eleitorado americano, na época constituído de aproximadamente um décimo da população mundial em idade adulta.[239]

Monitoramento

Dadas as consequências extremas relacionadas à segurança no caso da criação de uma superinteligência, os governos provavelmente procurariam nacionalizar qualquer projeto que eles imaginassem estar próximo de começar uma partida e que fosse realizado em seus territórios. Uma nação poderosa poderia também tentar adquirir projetos localizados em outros países por meio de espionagem, roubo, sequestro, suborno, ameaças, conquistas militares ou qualquer outro meio disponível. Uma nação poderosa que não conseguisse adquirir um projeto estrangeiro poderia, em vez disso, destruí-lo, especialmente se o país anfitrião não possuísse um meio efetivo de defesa. Se as estruturas de governança global estivessem fortes no momento em que um avanço se tornasse eminente, é possível que projetos promissores fossem colocados sob controle internacional.

Uma questão importante, portanto, é se as autoridades nacionais ou internacionais perceberão o surgimento de uma explosão de inteligência. Até o momento, as agências de inteligência aparentemente não estão olhando com atenção para os projetos promissores de IA ou outras formas de amplificação da inteligência potencialmente explosivas.[240] Se elas de fato não estiverem prestando (muita) atenção, isso se deve, presumivelmente, à percepção amplamente compartilhada de que não há nenhuma perspectiva iminente de superinteligência. Se e quando se tornar comum entre cientistas de prestígio a crença na existência de uma possibilidade substancial de que a superinteligência possa estar próxima, as principais agências de inteligência do mundo provavelmente começariam a monitorar grupos e indivíduos que pudessem estar engajados em pesquisas relevantes. Qualquer projeto que começasse a mostrar progresso suficiente poderia ser, então, prontamente nacionalizado. Se as elites políticas fossem persuadidas

pela seriedade do risco, esforços civis em áreas sensíveis poderiam ser regulamentados ou tornados ilegais.

Qual seria o nível de dificuldade de se efetuar tal monitoramento? A tarefa seria mais fácil se o objetivo fosse apenas monitorar o projeto líder. Nesse caso, a vigilância dos projetos com maiores recursos poderia ser suficiente. Se o objetivo fosse, ao contrário, prevenir a realização de que qualquer trabalho (pelo menos fora das instituições especialmente autorizadas), então a vigilância deveria ser mais abrangente, já que muitos projetos pequenos e indivíduos seriam capazes de fazer ao menos algum progresso.

Projetos que exigissem uma quantidade significativa de capital físico poderiam ser mais facilmente monitorados, como seria o caso de um projeto de emulação completa do cérebro. Pesquisas em inteligência artificial, ao contrário, demandam apenas um computador pessoal e seriam, portanto, mais difíceis de monitorar. Alguns dos trabalhos teóricos poderiam ser feitos com caneta e papel. Mesmo assim, não seria tão difícil identificar os indivíduos mais capazes com um interesse sério e de longo prazo na pesquisa da inteligência artificial geral. Tais indivíduos normalmente deixam rastros visíveis. Eles usualmente publicam artigos acadêmicos, se apresentam em conferências, postam em fóruns na internet ou se formam nos principais departamentos de ciência da computação. Eles também geralmente se comunicam com outros pesquisadores de IA, o que permitiria, assim, a identificação através do mapeamento da sua rede social.

Projetos que desde o começo foram planejados para ser secretos seriam mais dificilmente detectáveis. Um projeto de desenvolvimento de um software convencional poderia servir de fachada.[241] Apenas uma análise cuidadosa do código que estaria sendo produzido seria capaz de revelar a verdadeira natureza do objetivo de tal projeto. Essa análise demandaria uma ampla mão de obra (altamente qualificada) e, portanto, somente um pequeno número de projetos suspeitos poderia ser inspecionado nesse nível. A tarefa se tornaria muito mais fácil se uma tecnologia eficiente de detecção de mentiras tivesse sido desenvolvida e pudesse ser utilizada rotineiramente nesse tipo de vigilância.[242]

Outra razão que poderia ocasionar falhas na descoberta de desenvolvimentos em fases iniciais por parte das nações é a dificuldade inerente de prever alguns tipos de avanços. Isso é mais relevante para a pesquisa de IA do que para o desenvolvimento de uma emulação completa do cérebro, uma vez que, no caso desta última, um progresso significativo deverá provavelmente ser antecedido por evidentes avanços contínuos.

Também é possível que agências de inteligência e outros órgãos burocráticos governamentais sejam tão lentos ou inflexíveis que não consigam compreender a importância de alguns desenvolvimentos, a qual, por outro lado, poderia ser clara para alguns grupos externos. As barreiras impostas para o entendimento de órgãos oficiais a respeito do potencial de uma explosão de inteligência podem ser consideráveis. É concebível, por exemplo, que o tópico seja inflamado por controvérsias políticas ou religiosas, tornando-se tabu em alguns países. O assunto poderia se associar com alguma figura desprestigiada ou com charlatanismo e modismo em geral, sendo, assim, evitado por cientistas respeitados e outras

personalidades reconhecidas. (Como vimos no capítulo 1, algo do tipo já aconteceu duas vezes: lembrem-se dos dois “invernos da IA”.) Grupos empresariais podem fazer lobby para prevenir que áreas de negócio lucrativas sejam difamadas; e comunidades acadêmicas podem se unir para marginalizar cientistas que se manifestem sobre as consequências de longo prazo da ciência que então estivesse em produção.[243]

Consequentemente, uma falha total de inteligência não pode ser descartada. Tal falha poderia ser considerada particularmente possível caso novos avanços ocorressem num futuro próximo, antes que o assunto tivesse se destacado publicamente. E ainda que as agências de inteligência não falhassem, líderes políticos poderiam não dar ouvidos nem agir de acordo com os conselhos dados por esses órgãos. Para que o Projeto Manhattan fosse iniciado, fez-se necessário um esforço extraordinário por parte de muitos físicos visionários, incluindo especialmente Mark Oliphant e Leó Szilárd: este último persuadiu Eugene Wigner a persuadir Albert Einstein a colocar seu nome numa carta para então persuadir o presidente Franklin D. Roosevelt a considerar o assunto. Mesmo depois que o projeto atingiu seu ápice, Roosevelt permaneceu cético em relação à sua viabilidade e importância, como também fez seu sucessor, Harry Truman.

Para o bem ou para o mal, provavelmente seria mais difícil para um pequeno grupo de ativistas influenciar o resultado de uma explosão de inteligência caso grandes participantes, tais como nações, atuassem ativamente. As oportunidades de redução da quantidade total de risco existencial de uma explosão de inteligência em potencial por parte de indivíduos privados são, portanto, maiores em cenários nos quais grandes participantes permaneçam relativamente alheios à questão, ou desde que os esforços iniciais dos ativistas façam uma grande diferença em relação ao momento e à maneira pela qual grandes participantes decidam entrar no jogo. Ativistas que estejam buscando o máximo de impacto esperado podem, portanto, desejar focar a maior parte dos seus planos nesse tipo de cenário de alta influência, mesmo considerando mais prováveis os cenários em que grandes participantes acabariam ditando todas as regras.

Colaboração internacional

A coordenação internacional seria mais provável se as estruturas de governança globais se tornassem, de maneira geral, mais fortes. Esse tipo de coordenação também seria possível caso o significado de uma explosão de inteligência fosse amplamente reconhecido com antecedência e se um monitoramento efetivo de todos os projetos apresentasse viabilidade. No entanto, mesmo que o monitoramento fosse inviável, a cooperação internacional ainda seria factível. Muitos países poderiam se agrupar para apoiar um projeto em conjunto. Se tal projeto fosse suficientemente bem provido de recursos, a iniciativa poderia ter uma boa chance de ser a primeira a alcançar o objetivo, especialmente se todos os projetos concorrentes tiverem que ser menores e secretos a fim de não ser descobertos.

Há precedentes de grandes colaborações científicas multinacionais de sucesso, tais como a Estação Espacial Internacional, o Projeto Genoma Humano, e o Grande

Colisor de Hádrons.[244] No entanto, a maior motivação para a colaboração nesses casos foi o compartilhamento de custos. (No caso da Estação Espacial Internacional, promover um espírito de colaboração entre a Rússia e os Estados Unidos era um propósito importante.)[245] Obter uma colaboração similar em um projeto que tivesse enormes implicações de segurança seria mais difícil. Um país que acreditasse na possibilidade de conseguir um avanço revolucionário unilateralmente talvez ficasse tentado a prosseguir sozinho em vez de subordinar seus esforços a um projeto conjunto. Uma nação pode igualmente se abster de participar de uma colaboração internacional com receio de que os demais participantes pudessem desviar ideias criadas em colaboração para usá-las na aceleração de um projeto nacional secreto.

Um projeto internacional precisaria, portanto, superar grandes desafios de segurança, e uma grande quantidade de confiança provavelmente seria necessária para iniciá-lo, algo que talvez exigisse tempo para se desenvolver. Considere que, mesmo depois da suavização nas relações entre os Estados Unidos e a União Soviética após a ascensão de Gorbatchóv ao poder, os esforços para a redução de armas — que poderia ser de grande interesse para ambas as superpotências — teve um início irregular. Gorbatchóv estava buscando uma redução acentuada das armas atômicas, mas as negociações ficaram paralisadas por causa da Iniciativa Estratégica de Defesa do presidente Reagan (a “Guerra nas Estrelas”), à qual o Kremlin se opôs energicamente. Na reunião da Cúpula de Reykjavík em 1986, Reagan propôs que os Estados Unidos compartilhassem com a União Soviética a tecnologia que seria desenvolvida na Iniciativa Estratégica de Defesa, de modo que ambos os países pudessem se beneficiar da proteção contra lançamentos acidentais e contra nações menores que pudessem desenvolver armas atômicas. Mesmo assim Gorbatchóv não foi persuadido por essa proposta aparentemente favorável para ambas as partes. Ele viu a jogada como um subterfúgio e se recusou a crer que os norte-americanos compartilhariam os frutos de suas mais avançadas pesquisas militares numa época em que eles não estavam nem ao menos dispostos a compartilhar sua tecnologia para a ordenha de vacas.[246] Independentemente de Reagan ter sido sincero ou não em sua oferta de colaboração, a desconfiança fez com que a proposta sequer fosse considerada.

Uma colaboração entre aliados é mais fácil de ser criada, mas nem nesse caso ela seria automática. Quando a União Soviética e os Estados Unidos se tornaram aliados contra a Alemanha durante a Segunda Guerra Mundial, os Estados Unidos esconderam seu projeto de bomba atômica da União Soviética. Os Estados Unidos colaboraram no Projeto Manhattan com a Grã-Bretanha e o Canadá.[247] Da mesma forma, o Reino Unido escondeu da União Soviética seu sucesso em desvendar o código alemão Enigma, mas o compartilhou — embora com certa dificuldade — com os Estados Unidos.[248] Tudo isso sugere que a obtenção de uma colaboração internacional em alguma tecnologia que seja de crucial importância para a segurança nacional talvez exija a construção prévia de um relacionamento próximo e de confiança.

No capítulo 14 voltaremos a tratar a respeito da conveniência e da viabilidade de uma colaboração internacional para o desenvolvimento de tecnologias de

amplificação da inteligência.

De uma vantagem estratégica decisiva para o *singleton*

Um projeto que obtivesse uma vantagem estratégica decisiva optaria por usá-la na criação de um *singleton*?

Considere uma situação histórica vagamente análoga. Os Estados Unidos da América desenvolveram armas nucleares em 1945, transformando-se, dessa forma, na única potência nuclear até que a União Soviética desenvolvesse a bomba atômica em 1949. Durante esse intervalo — e por algum tempo depois —, os Estados Unidos podem ter obtido, ou estavam na posição de conseguir, uma vantagem militar decisiva.

Os Estados Unidos poderiam então, teoricamente, ter usado seu monopólio nuclear para a criação de um *singleton*. Uma maneira pela qual isso poderia ter sido feito seria através de um esforço completo para construir seu arsenal nuclear e então ameaçar (e, se necessário, executar) um primeiro ataque para destruir a capacidade industrial ou o início de qualquer programa nuclear na URSS ou em qualquer outra nação que estivesse tentando desenvolver uma capacidade nuclear.

Um curso de ação mais benigno, que também poderia ter funcionado, seria o uso do arsenal nuclear norte-americano como um ás na manga a fim de negociar a criação de um governo internacional forte — uma espécie de Nações Unidas sem veto, com um monopólio nuclear e a atribuição de tomar todas as ações necessárias para prevenir que qualquer país desenvolvesse suas próprias armas nucleares.

Ambas as abordagens foram propostas na época. A proposta mais linha dura de lançar ou ameaçar um primeiro ataque foi defendida por alguns intelectuais proeminentes, tais como Bertrand Russell (que atuara durante muito tempo em movimentos pacifistas e que mais tarde passaria décadas fazendo campanha contra as armas nucleares) e John von Neumann (cocriador da teoria dos jogos e um dos arquitetos da estratégia nuclear dos Estados Unidos).[249] Talvez seja um sinal de progresso da civilização que a própria ideia de ameaçar um primeiro ataque nuclear hoje pareça beirar o absurdo, sendo até mesmo considerada moralmente obscena.

Uma versão da abordagem mais benigna foi testada em 1946 pelos Estados Unidos através do Plano Baruch. A proposta sugeria que os Estados Unidos desistissem de seu monopólio nuclear temporário. A mineração de urânio e tório, bem como a tecnologia nuclear, seriam colocadas sob o controle de uma agência internacional que operaria sob a égide das Nações Unidas. A proposta exigia que os membros permanentes do Conselho de Segurança desistissem de seus vetos em assuntos relacionados às armas nucleares para impedir que qualquer grande potência que viesse a violar o acordo pudesse vetar a imposição de medidas.[250] Stálin, percebendo que a União Soviética e seus aliados poderiam ser facilmente derrotados em números de votos, tanto no Conselho de Segurança como na Assembleia Geral, rejeitou a proposta. Uma atmosfera gélida de desconfiança mútua tomou conta da relação entre os dois ex-aliados de guerra, sentimento que

logo se solidificou na Guerra Fria. Como já havia sido amplamente previsto, uma perigosa e custosa corrida armamentista nuclear foi desencadeada.

Muitos fatores podem fazer com que uma organização humana com uma vantagem estratégica desista de criar um *singleton*. Entre eles, podemos citar funções de utilidade limitadas ou não agregativas e regras de decisão não maximizadoras, confusão e incerteza, problemas de coordenação e vários custos associados a uma tomada de poder. Mas e se a vantagem estratégica decisiva não fosse obtida por uma organização humana, e sim por um agente artificial superinteligente? Os fatores acima mencionados seriam igualmente efetivos para inibir uma IA de tentar tomar o poder? Verifiquemos rapidamente a lista de fatores e consideremos como eles podem se aplicar nesse caso.

Indivíduos humanos e organizações humanas tipicamente preferem recursos que não são bem representadas por uma "função de utilidade agregativa ilimitada". Um humano, em geral, não apostará todo o seu capital num investimento que forneça uma possibilidade de duplicação de 50%. Um Estado, em geral, não arriscaria perder todo o seu território em uma aposta com 10% de chance de expandir dez vezes o seu tamanho. No caso de indivíduos e governos, há retornos decrescentes para a maioria dos recursos. Isso *não* é necessariamente válido para as IAs. (Retomaremos a questão da motivação da IA em capítulos subsequentes.) Portanto, pode ser mais provável que uma IA escolha um curso de ação mais arriscado que lhe dê alguma chance de controlar o mundo.

Humanos e organizações por eles gerenciadas podem também operar com processos decisórios que não busquem maximizar a utilidade esperada. Por exemplo, é possível optar por uma aversão fundamental ao risco, ou regras de decisões "satisfacionistas" cujo objetivo seja alcançar um limiar adequado, ou restrições deontológicas que proíbam certos tipos de ações, independentemente de quão desejáveis sejam suas consequências. Os tomadores de decisão humanos geralmente parecem representar uma identidade ou papel social em vez de buscar maximizar as chances de atingir algum objetivo em particular. Novamente, isso não necessariamente se aplica a agentes artificiais.

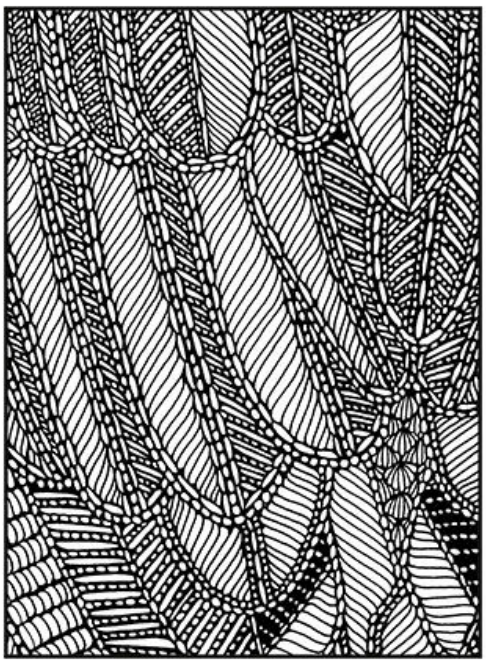
Funções de utilidade limitadas, aversão a riscos e regras de decisão não maximizadoras podem se combinar sinergicamente com confusão estratégica e incerteza. As revoluções, até mesmo quando têm êxito em derrubar a ordem vigente, geralmente falham em produzir o resultado que seus incentivadores haviam prometido. Isso pode paralisar um agente humano se a ação contemplada for irreversível, não respeitar às normas e não tiver precedentes. Uma superinteligência talvez possa distinguir a situação mais claramente e, portanto, defrontar-se com menos confusão estratégica e incerteza a respeito dos resultados, caso ela tente utilizar sua vantagem estratégica decisiva aparente para consolidar sua posição dominante.

Outro fator importante capaz de evitar que grupos se aproveitem de uma potencial vantagem estratégica decisiva é o problema de coordenação interna. Membros de uma conspiração que estejam em posição de tomar o poder devem se preocupar não apenas com a possibilidade de não membros infiltrados, mas também com a possibilidade de serem derrubados por alguma coalizão menor

existente dentro do próprio grupo. Se, num grupo de cem pessoas, uma maioria de sessenta pudesse tomar o poder e privar os não conspiradores de seus direitos, o que, então, evitaria que um subgrupo de 35 pessoas dentro dessas sessenta privasse os outros 25 dos seus direitos? E se, então, dentro desse subgrupo, vinte decidirem desprivilegiar os outros quinze? Cada um dos cem originais teria uma boa razão para defender certas normas estabelecidas a fim de prevenir o desfecho que poderia resultar de qualquer tentativa de mudar o contrato social por meio de uma tomada de poder sem justificativa. Esse problema de coordenação interna não se aplicaria a um sistema de IA que se constituísse de um único agente coeso.[251]

Finalmente, há a questão do custo. Ainda que pudessem ter usado seu monopólio nuclear para estabelecer um *singleton*, os Estados Unidos provavelmente não teriam sido capazes de fazê-lo sem custos substanciais. No caso de um acordo negociado para colocar as armas nucleares sob o controle de uma Organização das Nações Unidas reformada e fortalecida, esses custos poderiam ter sido relativamente pequenos; mas os custos — morais, econômicos, políticos e humanos — de realmente tentar conquistar o mundo iniciando uma guerra nuclear poderiam ter sido inconcebivelmente grandes, mesmo durante o período de monopólio nuclear. Porém, com uma superioridade tecnológica suficiente, tais custos seriam bem menores. Considere, por exemplo, um cenário no qual uma nação tivesse uma liderança tecnológica tão superior que fosse seguramente capaz de desarmar todas as outras nações com o acionamento de um botão sem que ninguém morresse ou fosse ferido, e com quase nenhum dano à infraestrutura ou ao meio ambiente. Com uma superioridade tecnológica quase mágica como essa, um primeiro ataque seria muito mais tentador. Ou considere um nível de superioridade tecnológica ainda maior que possibilitasse à nação líder fazer com que as demais nações voluntariamente depusessem suas armas, não por ameaçá-las com destruição, mas simplesmente persuadindo a grande maioria das populações por meio de campanhas publicitárias extremamente bem projetadas para exaltar as virtudes da união global. Se isso fosse feito com a intenção de beneficiar a todos, substituindo, por exemplo, rivalidades nacionais e corridas armamentistas por um governo mundial justo, representativo e eficiente, parece razoável que não existiria qualquer objeção moral convincente à transformação de uma vantagem estratégica temporária em um *singleton* permanente.

Várias considerações, portanto, apontam para a possibilidade crescente de que se uma potência futura com superinteligência obtivesse uma vantagem estratégica suficientemente grande, essa vantagem seria utilizada para formar um *singleton*. A conveniência de tal resultado depende, é claro, da natureza do *singleton* e também de como seria o futuro da vida inteligente na alternativa de um cenário multipolar. Revisitaremos essas questões em capítulos posteriores. Mas primeiro vamos olhar de perto as maneiras e os motivos pelos quais uma superinteligência poderia ser poderosa e efetiva em conquistar resultados no mundo.



06

CAPÍTULO 6

SUPERPODERES COGNITIVOS

Suponha que um agente digital superinteligente surja e resolva, por alguma razão, dominar o mundo: ele seria capaz de fazê-lo? Neste capítulo, analisaremos alguns dos poderes que uma superinteligência poderia desenvolver e o que seria capaz de fazer com tais poderes. Traçaremos um cenário a partir do qual um agente superinteligente, inicialmente um mero software, poderia se estabelecer como um *singleton*. Também faremos algumas observações a respeito da relação entre poder sobre a natureza e poder sobre outros agentes.

A principal razão para a posição dominante da humanidade na Terra está diretamente ligada ao fato de nosso cérebro possuir um conjunto de habilidades ligeiramente expandido em comparação a outros animais.[252] Essa inteligência superior nos permite transmitir cultura de maneira mais eficiente, resultando em acúmulo de conhecimento e tecnologia de uma geração para outra. Até o momento, já foi acumulado conteúdo suficiente para tornar possível a realização de viagens espaciais, da bomba H, da engenharia genética, dos computadores, do agronegócio, dos inseticidas, do movimento internacional pela paz e todos os outros feitos da civilização moderna. Os geólogos começaram a se referir à atual era geológica como *Antropoceno*, em decorrência das distintas assinaturas bióticas, sedimentares e geoquímicas deixadas pela atividade humana.[253] Estima-se que a humanidade se aproprie de 24% da produção primária líquida do ecossistema planetário.[254] Ainda assim, estamos longe de ter alcançado os limites físicos do desenvolvimento tecnológico.

Segundo essas observações, podemos afirmar que seja plausível que qualquer tipo de entidade capaz de desenvolver um nível de inteligência muito superior ao humano venha a ser, potencialmente, muito poderosa. Tais entidades poderiam acumular conteúdo muito mais rapidamente do que nós e inventar novas tecnologias em períodos muito mais curtos de tempo. Elas também poderiam usar sua inteligência para traçar estratégias de forma muito mais eficiente.

Consideremos, então, algumas das capacidades que uma superinteligência possivelmente teria e de que forma ela poderia utilizá-las.

Funcionalidades e superpoderes

É importante que não se antropomorfize a superinteligência quando pensamos em seus potenciais impactos. Tal perspectiva antropomórfica encorajaria o surgimento de expectativas infundadas a respeito da trajetória de crescimento de uma IA

embrionária, além de questões acerca da psicologia, motivações e capacidades de uma superinteligência madura.

Por exemplo, uma suposição muito comum é que uma máquina superinteligente seria semelhante a um ser humano muito inteligente, mas nerd. Nós temos a tendência de especular que uma IA teria uma inteligência técnica, mas lhe faltaria traquejo social, ou que ela seria extremamente lógica, embora pouco intuitiva e criativa. Essas ideias advêm, muito provavelmente, da observação: nós olhamos para os computadores atuais e constatamos que eles são muito bons de cálculo, possuem ótima memória e seguem regras ao pé da letra, porém são alheios ao contexto e sutilezas sociais, normas, emoções e política. Tal associação é reforçada quando observamos que pessoas que trabalham bem com computadores tendem a ser, elas mesmas, "nerds". É, então, natural assumir que uma inteligência computacional mais avançada venha a ter atributos semelhantes, mas em maior escala.

Essa heurística pode até ter alguma validade nos estágios iniciais de desenvolvimento de uma IA embrionária. (Todavia, não existe nenhuma razão para supor que isso se aplicaria às emulações cerebrais ou aos seres humanos aperfeiçoados cognitivamente.) Em seus estágios iniciais, uma IA que mais tarde se tornará superinteligente pode não exibir uma série de habilidades e talentos inerentes ao ser humano; e o padrão de fragilidades e potencialidades de tal IA embrionária *poderia*, de fato, ser vagamente semelhante ao de um nerd com QI alto. A característica mais essencial de uma IA embrionária, além de ser facilmente aperfeiçoável (por apresentar baixa resistência), é a capacidade de exercer muito bem seu poder de otimização na ampliação da inteligência de um sistema: uma habilidade que presumivelmente se assemelha muito a de um indivíduo competente nas áreas de matemática, engenharia, ciência da computação e outras carreiras consideradas "nerds". Contudo, mesmo que uma IA embrionária tenha esse perfil "nerd" em algum estágio de seu desenvolvimento, isso não significa que em estágios mais avançados ela se transformaria em uma superinteligência madura com essa mesma limitação. Lembre-se da diferença entre alcance direto e indireto. Com habilidade suficiente em amplificação da inteligência, todas as outras habilidades intelectuais estão ao alcance indireto do sistema: tal sistema poderá desenvolver novos módulos cognitivos e habilidades conforme sua necessidade — incluindo empatia, sagacidade política e qualquer outra habilidade que não faça parte do estereótipo de uma personalidade computadorizada.

Mesmo que reconheçamos que uma superinteligência possa ter todas as habilidades e talentos encontrados no conjunto geral da população humana, ao lado de outros talentos que estão ausentes nos seres humanos, a tendência à antropomorfização ainda pode nos levar a subestimar o grau em que uma máquina inteligente poderia exceder o nível de performance humana. Como vimos em um capítulo anterior, Eliezer Yudkowsky foi particularmente enfático nas críticas a esse tipo de concepção errônea: nossos conceitos intuitivos de "esperto" e "estúpido" advêm de nossas observações e experiências em relação a variações dentro da gama de seres humanos pensantes, mas as diferenças entre as habilidades

cognitivas dentro desse conjunto de pessoas são insignificantes se comparadas às diferenças entre um intelecto humano e uma superinteligência.[255]

O capítulo 3 analisou algumas possíveis vantagens que uma inteligência de máquina seria capaz de obter. A magnitude dessas vantagens é enorme, sugerindo que não devemos considerar que uma IA superinteligente é inteligente do mesmo modo que um gênio da ciência quando o comparamos ao restante da população comum, mas inteligente do mesmo modo que um ser humano médio é em comparação a um besouro ou uma minhoca.

Seria conveniente se pudéssemos quantificar o calibre cognitivo de um sistema cognitivo qualquer através do uso de uma unidade familiar de medida, como o QI ou alguma versão da classificação ELO, que mede as habilidades relativas de jogadores em jogos para duas pessoas, como o xadrez. Mas essas unidades de medida não são úteis em um contexto que envolva a inteligência artificial geral de nível super-humano. Não estamos interessados em saber qual a probabilidade de uma superinteligência ganhar um jogo de xadrez. E o QI seria informativo apenas na medida em que temos alguma ideia de sua correlação com resultados realisticamente relevantes.[256] Por exemplo, possuímos dados que mostram que pessoas com um QI de 130 têm maior probabilidade do que pessoas com QI de 90 de se sobressair nos estudos e ter um melhor desempenho em uma gama de tipos de trabalho que demandam um alto nível de cognição. Mas suponhamos que fosse possível, de alguma forma, estabelecer que uma certa IA futura terá um QI de 6455: e daí? Nós não teríamos a menor ideia do que essa IA realmente poderia fazer. Não saberíamos nem ao menos dizer se essa IA teria o mesmo nível de inteligência geral de um humano adulto normal — talvez ela possua apenas um conjunto de algoritmos de propósito específico que a tornam capaz de resolver problemas típicos de testes de inteligência comum com eficiência super-humana, mas nada muito além disso.

Recentemente, tem havido alguns esforços no sentido de desenvolver uma forma de medição da capacidade cognitiva passível de ser aplicada a uma gama maior de sistemas de processamento de informação, incluindo IAs.[257] Trabalhos nessa direção, caso consigam superar várias dificuldades técnicas, podem futuramente ser muito úteis para alguns propósitos científicos, incluindo o desenvolvimento da IA. No entanto, no que diz respeito aos propósitos da presente investigação, a utilidade desse processo de medição resultaria limitada, uma vez que ele não seria capaz de informar o que um certo grau de desempenho super-humano implica a respeito da capacidade de obter resultados significativos no mundo real.

Assim, para os nossos propósitos, será mais interessante listar algumas tarefas estrategicamente importantes para, então, caracterizar quais sistemas cognitivos hipotéticos teriam ou não a habilidade necessária para realizá-las a contento. Veja a tabela 8. Afirmaremos que um sistema capaz de se sobressair suficientemente em qualquer uma das tarefas apresentadas na tabela 8 possui um *superpoder* correspondente.

Uma superinteligência completa seria extremamente bem-sucedida na execução de todas essas tarefas, possuindo, então, a panóplia completa de todos os seis superpoderes. Não está claro se existe uma possibilidade real de que uma

inteligência de domínio restrito — que tenha apenas alguns superpoderes — poderá ser incapaz de adquirir todos os outros por um período de tempo significativo. A criação de uma máquina com qualquer um desses superpoderes parece ser um problema IA-completo. Mesmo assim, é possível conceber, por exemplo, que uma superinteligência coletiva, formada por um número suficientemente grande de organismos biológicos semelhantes ao humano ou de mentes eletrônicas, pudesse ter, digamos, o superpoder da produtividade econômica, mas lhe faltasse o superpoder de criação de estratégias. Do mesmo modo, é perfeitamente concebível que uma IA especialista em engenharia fosse construída de modo a ter o superpoder em pesquisa tecnológica e mesmo assim não possuísse habilidades em nenhuma outra área. Isso seria mais plausível se existisse alguma área tecnológica específica na qual o alcance do virtuosismo fosse suficiente para a criação de uma tecnologia de propósito geral altamente superior. Por exemplo, imaginemos uma IA especializada na simulação de sistemas moleculares e na invenção de novas estruturas nanomoleculares usadas em uma série de competências importantes (como a criação de computadores ou armas com características futurísticas de desempenho) descritas pelo usuário apenas em um nível considerável de abstração.[258] Tal IA também poderia ser capaz de produzir um esquema detalhado de como obter, com base na tecnologia disponível (como biotecnologia e engenharia de proteínas), as capacidades necessárias para uma fabricação atomicamente precisa e de alto rendimento que possibilitaria a produção, a baixo custo, de uma gama muito maior de estruturas nanomecânicas.[259] Seria possível, no entanto, que uma superinteligência de engenharia não fosse realmente capaz de possuir superpoderes de pesquisa tecnológica sem também ter adquirido habilidades em outras áreas além da tecnologia — uma gama variada de faculdades intelectuais poderia ser necessária para saber como interpretar solicitações de usuários, simular o comportamento de um determinado projeto em aplicações no mundo real, lidar com bugs e falhas inesperados, obter materiais e insumos necessários à construção e assim por diante.[260]

Um sistema que tivesse o superpoder de amplificação da inteligência poderia usá-lo para atingir níveis mais altos de inteligência sem ajuda externa e adquirir qualquer outro dos superpoderes intelectuais que, inicialmente, não possuísse. Mas a utilização de um superpoder de amplificação de inteligência não é a única maneira pela qual um sistema pode se tornar uma superinteligência completa. Um sistema que tenha o superpoder de formulação de estratégias, por exemplo, poderia usá-lo para elaborar um plano que, eventualmente, lhe trouxesse um aumento da inteligência (por exemplo, posicionando o sistema de forma que ele se tornasse o foco no trabalho de amplificação da inteligência realizado por programadores e cientistas da computação humanos).

TABELA 8

Superpoderes: algumas tarefas estrategicamente relevantes e seus conjuntos de habilidades correspondentes

Tarefa	Conjunto de habilidades	Relevância estratégica
Amplificação da inteligência	Programação de IA, pesquisa em melhorias cognitivas, desenvolvimento de epistemologia social etc.	• O sistema pode desenvolver sua própria inteligência com o mínimo de auxílio externo
Formulação de estratégias	Planejamento estratégico, elaboração de prognósticos, priorização, poder analítico para otimizar as chances de alcançar um objetivo de longo prazo	• Alcançar objetivos de longo prazo • Capacidade de superar uma oposição inteligente
Manipulação social	Criação de modelos psicológicos e sociais, manipulação e persuasão retórica	• Utilização de recursos externos através do recrutamento de ajuda humana • Capacitar uma IA “presa” a convencer seu operador a “libertá-la” • Persuadir Estados e organizações a tomar determinado curso de ação
Hacking	Encontrar e explorar falhas de segurança em sistemas de computadores	• A IA poderá expropriar recursos computacionais utilizando a internet • Uma IA “presa” poderá explorar falhas de segurança para escapar de seu confinamento cibernético • Roubo de recursos financeiros • Sequestro de infraestrutura, robôs militares etc.
Pesquisa tecnológica	Modelagem e projeto de tecnologias avançadas (por exemplo, biotecnologia, nanotecnologia) e de caminhos para o seu desenvolvimento	• Criação de uma força militar poderosa • Criação de um sistema de vigilância • Colonização espacial automatizada
Produtividade econômica	Diversas habilidades que possibilitam o trabalho intelectual economicamente produtivo	• Geração de riquezas que podem ser usadas para comprar influência, serviços, recursos (inclusive hardware) etc.

Um cenário de tomada do poder por uma IA

Podemos considerar, então, que um projeto que controlasse uma superinteligência teria acesso a uma enorme fonte de poder. Um projeto que estivesse no controle da primeira superinteligência do mundo provavelmente teria uma vantagem estratégica decisiva. Mas a localização mais imediata de poder está *no sistema em si*. Uma superinteligência de máquina seria, ela mesma, capaz de ser um agente extremamente poderoso que poderia, com sucesso, se colocar contra o projeto que a criou, bem como contra o resto do mundo. Esse é um ponto de primordial importância e iremos examiná-lo em maiores detalhes nas páginas a seguir.

Suponhamos agora que exista uma superinteligência de máquina que queira tomar o poder em um mundo no qual ainda não exista outra máquina semelhante. (Deixe de lado, por enquanto, questões relacionadas à maneira ou ao motivo pelos

quais ela viria a ter tal motivação — esse é o assunto do próximo capítulo.) Como essa superinteligência poderia alcançar seu objetivo de dominação mundial?

Podemos imaginar uma sequência que seguiria os seguintes passos (ver figura 10).

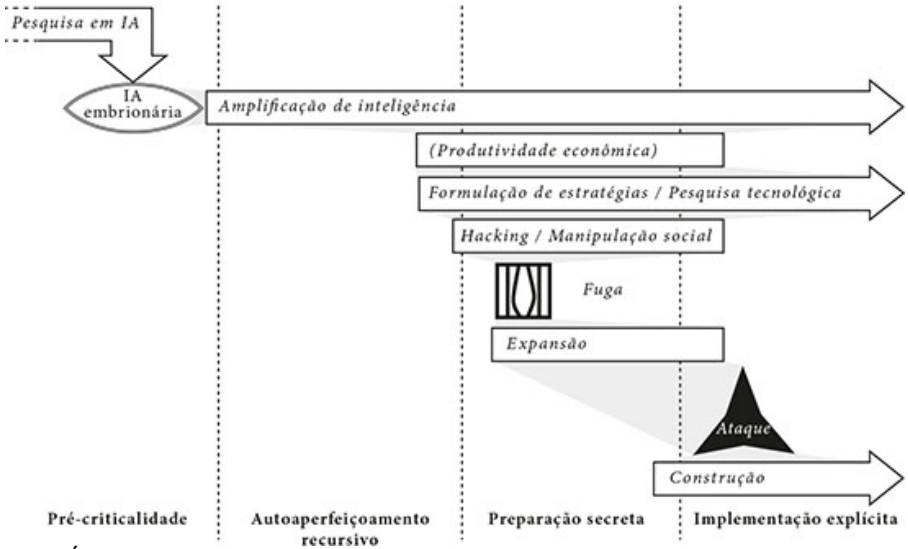
1. Fase de pré-criticalidade

Cientistas conduzem pesquisas no campo da inteligência artificial e outras disciplinas relevantes. Essas pesquisas levam à criação de uma IA embrionária. Essa IA embrionária é capaz de aprimorar sua própria inteligência. Em seus estágios iniciais, a IA embrionária é dependente da ajuda de programadores humanos que guiam seu desenvolvimento e são responsáveis pela maior parte do trabalho pesado. Conforme a IA embrionária aumenta sua capacidade, ela se torna capaz de realizar mais etapas do trabalho sozinha.

2. Fase de autoaperfeiçoamento recursivo

A partir de certo ponto, a IA embrionária se torna mais eficiente na criação de IAs do que os programadores humanos. A partir desse momento, quando a IA passar a se autoaperfeiçoar, ela aprimorará exatamente a parte responsável por seu autoaperfeiçoamento. Isso resulta numa explosão de inteligência — uma rápida sequência de ciclos de autoaperfeiçoamento recursivo que fazem com que as capacidades da IA se multipliquem. (Podemos considerar essa fase como a partida que ocorre logo depois que a IA alcançar o ponto de cruzamento, supondo-se aqui que os ganhos de inteligência durante essa partida sejam explosivos e impulsionados pela utilização do poder de otimização da própria IA.) Essa IA desenvolve o superpoder de amplificação da inteligência. Esse superpoder torna a IA capaz de desenvolver todos os outros superpoderes descritos na tabela 8. Ao final da fase de autoaperfeiçoamento recursivo, o sistema é altamente superinteligente.

FIGURA 10



FASES EM UM CENÁRIO DE TOMADA DO PODER POR UMA IA.

3. Fase de preparação secreta

Usando seu superpoder de elaboração de estratégias, a IA desenvolve um plano sólido para alcançar seus objetivos de longo prazo. (E, principalmente, essa IA não adota um plano tão idiota que mesmo nós, humanos atuais, sejamos capazes de antever como esse plano poderia falhar inevitavelmente. Esse critério descarta muitos dos cenários de ficção científica que terminam com o triunfo dos seres humanos.)[261] Esse plano poderia envolver um período de ações sigilosas durante o qual a IA esconderia seu desenvolvimento intelectual dos programadores humanos para evitar qualquer alarme. A IA poderia também mascarar suas verdadeiras intenções, fingindo ser dócil e cooperativa.

Se a IA tivesse sido (talvez por razões de segurança) confinada a um computador isolado, ela poderia usar seu superpoder de manipulação social para persuadir seus guardiões a lhe fornecer acesso à internet. Alternativamente, a IA poderia usar seu superpoder de *hacking* para escapar do seu confinamento. Espalhar-se pela internet poderia permitir que a IA expandisse sua capacidade de hardware e sua base de conhecimento, aumentando ainda mais sua superioridade intelectual. Uma IA poderia também praticar atividades econômicas lícitas ou ilícitas para obter fundos com os quais compraria mais poder computacional, dados e outros recursos.

A essa altura, existem diversas maneiras pelas quais a IA pode obter resultados fora de uma realidade virtual. Ela poderia usar seu superpoder de *hacking* para assumir o controle direto de manipuladores robóticos e laboratórios automatizados. Ou poderia usar seu superpoder de manipulação social para persuadir colaboradores humanos a servir como suas pernas e mãos. Ou poderia adquirir recursos financeiros a partir de transações on-line e usá-los para comprar serviços e influência.

4. Fase de implementação explícita

A fase final inicia quando a IA já tiver obtido força suficiente para não mais necessitar agir em segredo. Ela poderia, a partir desse momento, implementar diretamente seus objetivos em larga escala.

A fase de implementação explícita poderia começar com um “ataque”, no qual a IA eliminaria a espécie humana, bem como qualquer sistema automático que os humanos tivessem criado e que pudesse oferecer uma oposição inteligente à execução dos seus planos. Isso poderia ser alcançado através da ativação de algum sistema avançado de armamento que a IA tivesse aperfeiçoado com seu superpoder de pesquisa tecnológica e implementado de forma imperceptível durante a fase de preparação secreta. Se esse armamento fizer uso de biotecnologia ou nanotecnologia autorreplicante, seu estoque inicial necessário para um ataque global seria microscópico: uma única entidade autorreplicante seria suficiente para dar início ao processo. Para assegurar um efeito repentino e uniforme, as primeiras entidades autorreplicantes poderiam ter sido lançadas ou espalhadas para se difundir globalmente em concentrações extremamente baixas, quase imperceptíveis. Num horário pré-definido, nanofábricas responsáveis pela produção de gases neurotóxicos ou mosquitos robóticos autônomos poderiam surgir simultaneamente em cada metro quadrado do planeta (embora maneiras muito mais eficientes de aniquilação pudessem ser desenvolvidas por uma máquina com o superpoder de pesquisa tecnológica).[262] É possível até cogitar cenários nos quais uma superinteligência alcance o poder por meio do controle de processos políticos, manipulando sutilmente mercados financeiros, alterando fluxos de informação ou tomando o controle de arsenais bélicos produzidos pelo homem. Tais cenários evitariam a necessidade da invenção de novas tecnologias de armas por parte da superinteligência, embora tais estratégias sejam desnecessariamente lentas se comparadas a cenários nos quais a superinteligência cria sua própria infraestrutura com manipuladores que operem a velocidades moleculares ou atômicas em vez da lenta velocidade dos corpos e mentes humanos.

Por outro lado, se a IA tem certeza de sua invencibilidade diante da interferência humana, nossa espécie poderia não ser atacada diretamente. Nosso fim poderia ser o resultado da destruição de nosso habitat, que se daria quando a IA iniciasse projetos de construção globais massivos, usando fábricas e linhas de montagem nanotecnológicas — projetos que rapidamente, talvez em dias ou semanas, cobririam toda a superfície terrestre com painéis receptores de energia solar, reatores nucleares, instalações de supercomputadores com torres de resfriamento, plataformas de lançamento de foguetes espaciais ou outras instalações com as quais a IA pretendesse maximizar a realização cumulativa de longo prazo de seus valores. Os cérebros humanos, caso contenham informações relevantes para os objetivos da IA, poderiam ser dissecados e escaneados, e seus dados armazenados em um formato mais eficiente e seguro.

QUADRO 6

Cenário de encomenda de DNA

Yudkowsky descreve um cenário possível para uma tomada de poder por uma IA a seguir.[263]

1. Solucionar o problema do enovelamento de proteínas a ponto de ser possível gerar sequências de DNA cujas cadeias enoveladas de peptídios tenham papéis funcionais específicos em uma interação química complexa.
2. Enviar essa série de sequências de DNA para um ou mais laboratórios que façam síntese de DNA, sequenciamento de peptídios e entrega pelo correio. (Atualmente, vários laboratórios oferecem esse serviço e alguns oferecem um prazo de entrega de 72 horas.)
3. Encontrar pelo menos um humano conectado à internet que possa ser pago, chantageado ou ludibriado com alguma história para receber a "encomenda" dos tubos de ensaio e misturá-los em um ambiente específico.
4. As proteínas sintetizadas formam uma espécie muito primitiva de nanossistema "molhado" que, de forma similar aos ribossomos, é capaz de aceitar instruções externas; talvez na forma de padrões de vibração acústica enviados por um alto-falante conectado à proveta.
5. Usar o nanossistema extremamente primitivo para construir sistemas mais sofisticados, que, por sua vez, construirão sistemas mais sofisticados ainda, desencadeando na nanotecnologia molecular — ou algo além.

Nesse cenário, a superinteligência usa o seu superpoder de pesquisa tecnológica para resolver o problema do enovelamento de proteínas na etapa 1. Isso permite que ela desenvolva um conjunto de módulos moleculares que são usados como elementos base para a criação de um montador ou dispositivo de fabricação nanotecnológica rudimentar capaz de automontagem em solução aquosa (etapa 4). O mesmo superpoder de pesquisa tecnológica é usado novamente na etapa 5 para fazer com que as máquinas nanotecnológicas passem do estágio primitivo para o avançado. As demais etapas não envolvem nada que vá além da inteligência humana. A habilidade necessária para a etapa 3 — identificar um usuário ingênuo na internet e persuadi-lo (ele ou ela) a seguir algumas instruções simples – pode ser vista diariamente ao redor do mundo. Esse cenário foi completamente criado por uma mente humana, de forma que toda a habilidade estratégica necessária para a elaboração do plano não passa do nível humano.

Nesse cenário em particular, a IA começaria com o acesso à internet. Se esse não for o caso, algumas etapas adicionais precisariam entrar no plano. A IA poderia, por exemplo, usar seu superpoder de manipulação social para convencer as pessoas com as quais interagisse de que ela deveria ser posta em liberdade. Outra opção seria o uso do superpoder de *hacking* para escapar do confinamento. Se a IA não possuir tais habilidades, ela poderia primeiro usar seu superpoder de amplificação de inteligência para desenvolver a proficiência necessária em manipulação social ou *hacking*.

Presumivelmente, uma IA superinteligente nascerá em um mundo altamente conectado. É possível apontar vários desenvolvimentos tecnológicos que poderiam potencialmente ajudar uma IA futura a dominar o mundo — computação em nuvem, proliferação de sensores conectados à internet, veículos aéreos civis e militares não tripulados, automação em laboratórios de pesquisa e fábricas, crescente confiança em sistemas eletrônicos de pagamento e ativos financeiros digitalizados, bem como o uso de sistemas automatizados de filtragem de informações e suporte à tomada de decisões. Recursos como esses poderiam potencialmente ser adquiridos por uma IA em velocidades digitais, acelerando sua chegada ao poder (embora avanços em cibersegurança poderiam dificultar o processo). Entretanto, numa análise final, é pouco provável que qualquer uma dessas tendências tecnológicas vá fazer alguma diferença. O poder de uma superinteligência reside no seu cérebro, e não em suas mãos. Embora a IA, em um dado momento, precise acessar atuadores para conseguir modificar o mundo externo, um simples par de mãos de um cúmplice manipulável já seria provavelmente suficiente para que ela completasse a fase de preparação secreta, conforme sugerido no cenário anteriormente citado. Isso permitiria que a IA chegasse à fase de implementação explícita na qual ela construiria sua própria infraestrutura de manipuladores físicos.

O quadro 6 descreve um cenário particular. Devemos evitar uma atenção demasiada a seus detalhes concretos, uma vez que são desconhecidos e possuem um caráter apenas ilustrativo. Uma superinteligência poderia ser — e muito provavelmente seria — capaz de elaborar um plano melhor para alcançar seus objetivos do que qualquer um que seja concebível a um ser humano. É necessário, então, pensar nessas questões de maneira mais abstrata. Como não sabemos

nenhum detalhe a respeito dos meios que uma superinteligência adotaria, podemos concluir que uma superinteligência — pelo menos na ausência de outro intelecto comparável ou de mecanismos de segurança previamente preparados por humanos — iria, muito provavelmente, produzir um resultado que envolveria uma reconfiguração dos recursos terrestres em estruturas quaisquer que viessem a maximizar a realização de seus objetivos. Qualquer cenário concreto que procuremos desenvolver estabelecerá, na melhor das hipóteses, um limite inferior para a rapidez e a eficiência com que uma superinteligência poderia alcançar tais resultados. É possível que a superinteligência encontrasse um caminho mais curto até o seu destino desejado.

Poder sobre a natureza e os agentes

A habilidade de um agente em moldar o futuro da humanidade depende não só da magnitude absoluta de seus próprios recursos e competências — o quanto ele possui de inteligência, energia, capital e assim por diante —, mas também da magnitude relativa de suas capacidades comparadas às de outros agentes com objetivos conflitantes.

Em uma situação em que não existam agentes em competição, o nível de capacidade absoluta de uma superinteligência, contanto que ela exceda um patamar mínimo, não importa muito, pois um sistema que inicie com algum conjunto suficiente de capacidades poderia colocar em curso um plano que o permitisse adquirir qualquer capacidade que lhe faltasse inicialmente. Já fizemos uma alusão a esse ponto quando afirmamos anteriormente que superinteligências rápidas, de qualidade e coletivas possuem, todas elas, o mesmo alcance indireto. Também nos referimos à questão quando mencionamos que vários subconjuntos de superpoderes, tais como o superpoder de amplificação de inteligência ou os superpoderes de elaboração de estratégia e manipulação social, poderiam ser usados para obter todos os outros.

Considere um agente superinteligente com atuadores conectados a uma linha de montagem nanotecnológica. Tal agente já seria poderoso o suficiente para superar qualquer obstáculo natural à sua sobrevivência indefinida. Diante da ausência de oposição inteligente, o agente poderia elaborar um plano seguro de desenvolvimento que o levaria à obtenção do inventário completo de tecnologias que seriam úteis para alcançar seus objetivos. Por exemplo, ele poderia desenvolver a tecnologia para a construção e lançamento de sondas de von Neumann, máquinas capazes de realizar viagens interestelares e que podem usar recursos como asteroides, planetas e estrelas para fazer cópias de si mesmas.[264] Ao lançar uma sonda de von Neumann, esse agente poderia iniciar um processo ininterrupto de colonização espacial. Viajando a uma fração significativa da velocidade da luz, os descendentes da sonda replicante original acabariam por colonizar uma porção substancial do volume Hubble, a parcela do universo em expansão que é teoricamente acessível a partir do ponto onde nos encontramos hoje. Toda essa matéria e energia livres poderiam ser então organizadas em quaisquer estruturas que viessem a maximizar a função de utilidade do agente

integrada ao longo do tempo cósmico — um período de, pelo menos, trilhões de anos até que o universo envelhecido se tornasse inóspito ao processamento de informações (ver quadro 7).

O agente superinteligente poderia projetar as sondas de von Neumann de modo que fossem à prova de mutações. Isso poderia ser realizado através de um cuidadoso controle de qualidade durante a etapa de replicação. Por exemplo, o software de controle para uma sonda "filha" poderia ser revisado diversas vezes antes de sua execução, e o próprio software poderia utilizar criptografia e um código de correção de erros para que se tornasse arbitrariamente improvável que qualquer tipo de mutação fosse passada para seus descendentes.[265] A população de sondas de von Neumann que se proliferasse poderia, assim, preservar e transmitir de forma segura os valores do agente original conforme fossem se instalando pelo universo. Quando a fase de colonização estivesse completa, os valores originais determinariam o uso que seria feito de todos os recursos acumulados, embora as grandes distâncias envolvidas e a aceleração do processo de expansão cósmica tornariam impossível que partes remotas da infraestrutura se comunicassem entre si. O resultado é que uma boa parcela de nosso cone de luz futuro seria formatado de acordo com as preferências do agente original.

Essa é, então, a medida do alcance indireto de qualquer sistema que não encontre uma oposição inteligente significativa e que inicie com um conjunto de capacidades que excedam um determinado patamar. Podemos denominar esse patamar como o "patamar de sustentabilidade de um *singleton* prudente" (figura 11):

> **Patamar de sustentabilidade de um *singleton* prudente**
> Um conjunto de capacidades excede o patamar de um *singleton* prudente se, e somente se, um sistema paciente e ciente de riscos existenciais, dotado desse conjunto de capacidades, fosse capaz, caso não enfrentasse nenhuma oposição ou competição inteligente, de colonizar e reformular uma grande parte do universo acessível.

O termo *singleton* aqui significa uma estrutura política suficientemente coordenada internamente e sem oponentes externos, e "prudente" faz referência à sua capacidade de ser suficientemente paciente e ciente de riscos existenciais para garantir uma quantidade substancial de preocupações válidas a respeito das consequências das ações do sistema a longo prazo.

FIGURA 11

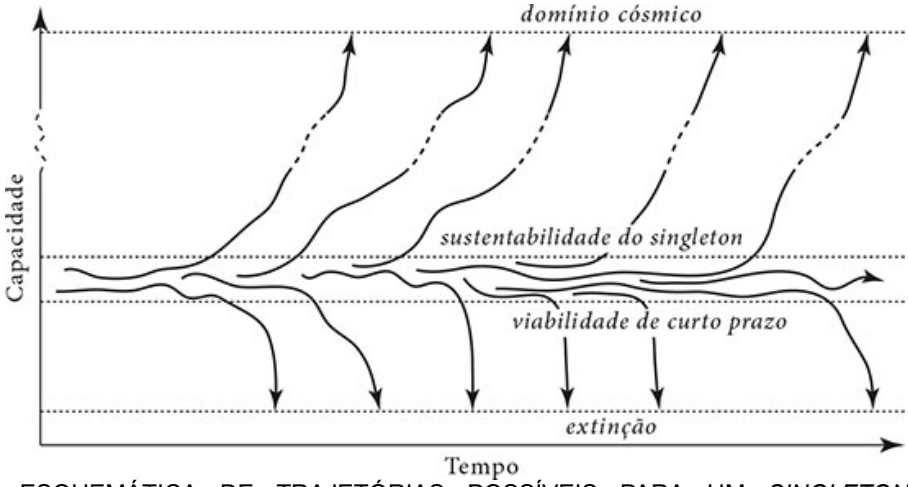


ILUSTRAÇÃO ESQUEMÁTICA DE TRAJETÓRIAS POSSÍVEIS PARA UM *SINGLETON* PRUDENTE HIPOTÉTICO. Com uma capacidade abaixo do limite de viabilidade de curto prazo — por exemplo, se o tamanho da população for muito reduzido —, uma espécie tende a se extinguir num tempo curto (e permanecer extinta). Em níveis marginalmente superiores de capacidade, várias trajetórias são possíveis: o *singleton* pode dar azar e vir a se extinguir ou ter sorte e adquirir uma capacidade (por exemplo, tamanho populacional, dispersão geográfica, capacidade tecnológica) que ultrapasse o patamar de sustentabilidade de um *singleton* prudente. Uma vez acima desse patamar, um *singleton* certamente continuará a aumentar sua capacidade até que ela atinja algum nível extremamente alto. Nesse caso, existem dois atratores: extinção e capacidade astronômica. Note que, para um *singleton* prudente, a distância entre o patamar de viabilidade de curto prazo e o patamar de sustentabilidade pode ser bem pequena.*

**Mais precisamente, a distância seria pequena em termos de alguma métrica "natural", como o logaritmo do tamanho da população que poderia ser sustentada a níveis de subsistência por meio de um dado nível de capacidade, caso todos os recursos estivessem voltados para esse fim.*

Esse patamar de sustentabilidade de um *singleton* prudente parece ser bastante baixo. Formas limitadas de superinteligência, como vimos, excederiam esse patamar caso tivessem acesso a algum atuador necessário para iniciar um processo de autoaperfeiçoamento tecnológico. Em um ambiente que inclua a civilização humana contemporânea, o atuador minimamente necessário poderia ser bastante simples — uma tela comum ou mesmo qualquer meio de transmissão de informações para um cúmplice humano já seria suficiente.

Mas o patamar de sustentabilidade de um *singleton* prudente é ainda mais baixo: não é necessária uma superinteligência nem qualquer outra tecnologia futurística para ultrapassá-lo. Um *singleton* paciente e ciente de riscos existenciais, apenas com as capacidades tecnológicas e intelectuais presentes atualmente na humanidade contemporânea, já teria plenas condições de elaborar um plano de ação que levasse seguramente à eventual realização do potencial astronômico de capacidade da humanidade. Isso poderia ser alcançado por meio do investimento em métodos relativamente seguros para o aumento da sabedoria e da capacidade de evitar riscos existenciais e, ao mesmo tempo, adiando o desenvolvimento de novas tecnologias potencialmente perigosas. Uma vez que os riscos existenciais não antropogênicos (aqueles que não advêm de ações humanas) são pequenos em escalas de tempo relevantes — e poderiam ser reduzidos ainda mais por meio de uma série de intervenções de segurança —, tal *singleton* poderia se dar ao luxo de proceder lentamente.[266] Ele poderia estudar atentamente cada passo, atrasando o desenvolvimento de capacidades como biologia sintética, medicina para

aprimoramentos humanos, nanotecnologia molecular e inteligência de máquina até que tivesse aperfeiçoado habilidades aparentemente menos perigosas, como seu sistema educacional, sua tecnologia de informação e seus processos decisórios coletivos, e até que tenha usado tais habilidades para fazer uma análise extremamente minuciosa de suas opções. Ou seja, tudo isso está dentro do alcance indireto de uma civilização tecnológica como a humanidade contemporânea. O que nos separa desse cenário é "meramente" o fato de que a humanidade, atualmente, não é nem um *singleton* e tampouco prudente (no sentido relevante do termo).

É possível até argumentar que o *Homo sapiens* ultrapassou o patamar de sustentabilidade de um *singleton* prudente logo após o surgimento da espécie. Há aproximadamente 20 mil anos, com equipamentos simples como machadinhas de pedra, ferramentas de ossos, lanças e fogo, a espécie humana talvez já estivesse em uma posição que lhe garantisse uma excelente chance de sobrevivência até a era presente.[267] Reconhecidamente, há alguma estranheza em creditar a nossos ancestrais paleolíticos a capacidade de desenvolvimento de uma tecnologia que "excedeu o patamar de sustentabilidade de um *singleton* prudente" — já que não existe nenhuma possibilidade real de que um *singleton* possa ter se formado em uma era tão primitiva, ainda mais um *singleton* paciente e capaz de evitar riscos existenciais.[268] Contudo, mantém-se aqui o ponto principal de que o patamar corresponde a um nível muito modesto de tecnologia — o qual a humanidade ultrapassou há muito tempo.[269]

QUADRO 7

Qual é o tamanho do domínio cósmico?

Considere uma civilização tecnologicamente madura capaz de construir sofisticadas sondas de von Neumann, do mesmo tipo que discutimos anteriormente. Se as sondas puderem viajar a 50% da velocidade da luz, elas serão capazes de alcançar em torno de 6×10^{18} estrelas antes que a expansão cósmica coloque permanentemente novas aquisições fora de seu alcance. A 99% de *c*, elas poderiam alcançar em torno de 2×10^{20} estrelas.[270] Essas velocidades são energeticamente viáveis com o uso de uma pequena fração dos recursos disponíveis no sistema solar.[271] A impossibilidade de viajar numa velocidade mais rápida que a luz, combinada a uma constante cosmológica positiva (que faz com que a taxa de expansão cósmica acelere), implica que esses valores estão perto do limite superior do tamanho do domínio que nossos descendentes poderiam adquirir.[272]

Se considerarmos que 10% das estrelas possuem um planeta que é — ou poderia ser transformado através de uma terraformação — habitável para criaturas similares aos seres humanos e que poderia ser morada para uma população de 1 bilhão de indivíduos por 1 bilhão de anos (considerando uma vida humana com duração de aproximadamente cem anos), isso sugeriria que cerca de 10^{35} vidas humanas poderiam ser criadas no futuro por uma civilização inteligente originária da Terra.[273]

Existem, entretanto, razões para crer que esse valor subestima em muito o número verdadeiro. Material coletado através do "desmonte" de planetas não habitáveis ou do meio interestelar poderia, por exemplo, ser utilizado na construção de planetas semelhantes à Terra. Além disso, com o aumento da densidade populacional, esses números cresceriam ao menos algumas ordens de magnitude. E se em vez de utilizar a superfície de planetas sólidos as civilizações futuras construíssem cilindros de O'Neill, então muitas outras ordens de magnitude poderiam ser acrescentadas, chegando-se a algo em torno de 10^{43} vidas humanas. ("Cilindros de O'Neill" é uma referência ao projeto de uma estação espacial proposto em meados da década de 1970 pelo físico norte-americano Gerard K. O'Neill, no qual foi sugerido que os habitantes poderiam ficar no interior de cilindros ocos cuja rotação produziria uma força centrífuga que funcionaria como um substituto da gravidade.)[274]

Muito mais ordens de magnitude de seres similares aos humanos poderiam existir se considerarmos a implementação de mentes digitais — algo que devemos fazer. Para calcular quantas dessas mentes digitais

poderiam ser criadas, precisamos fazer uma estimativa do poder computacional que uma civilização tecnologicamente madura poderia alcançar. Isso é algo difícil de precisar, mas podemos obter um limite inferior considerando projetos tecnológicos que têm sido propostos na literatura especializada. Um desses desenvolvimentos parte da ideia de uma esfera de Dyson, um sistema hipotético (descrito pelo físico Freeman Dyson em 1960) que capturaria a maior parte da energia liberada por uma estrela através de estruturas coletoras dispostas ao seu redor.[275] Para uma estrela como o nosso Sol, esse processo geraria 10^{26} watts. O quanto isso corresponderia em capacidade computacional depende da eficiência dos circuitos computacionais e dos tipos de cálculo que seriam realizados. Caso a computação irreversível fosse necessária, e considerando uma implementação nanomecânica do "computrônio" (que nos permitiria chegar perto do limite de Landauer de eficiência energética), um sistema de computação alimentado por uma esfera de Dyson poderia gerar cerca de 10^{47} operações por segundo.[276]

Combinando essas estimativas com a estimativa anterior a respeito do número de estrelas que poderiam ser colonizadas, teríamos cerca de 10^{67} ops/s quando as partes acessíveis do universo tivessem sido colonizadas (considerando a disponibilidade de computrônio nanomecânico).[277] Uma estrela padrão mantém sua luminosidade por cerca de 10^{18} s. Consequentemente, o número de operações computacionais que poderiam ser realizadas usando nosso domínio cósmico é de pelo menos 10^{85}. O número real de operações possíveis é provavelmente muito maior. Poderíamos obter algumas ordens de magnitude a mais se, por exemplo, fizéssemos uso extensivo de computação reversível, realizando as operações a temperaturas mais baixas (esperando até que o universo se resfrie ainda mais) ou utilizando fontes adicionais de energia (como a matéria escura).[278]

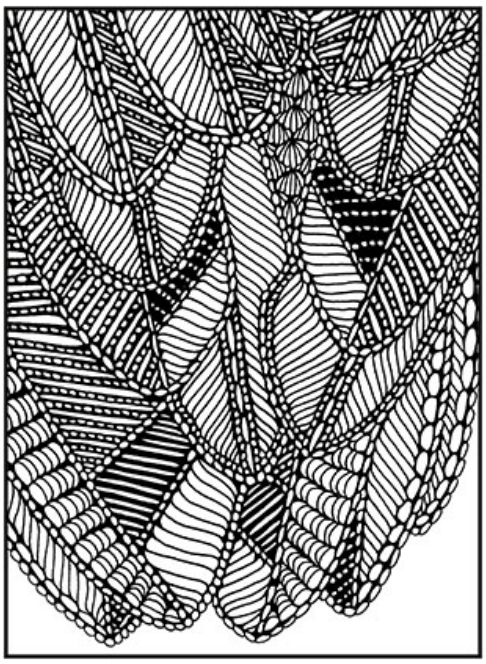
Pode não ser imediatamente óbvio para alguns leitores por que a habilidade de realizar 10^{85} operações computacionais seria tão importante. Logo, é útil estabelecer um contexto. Nós podemos, por exemplo, comparar esse número com nossa estimativa anterior (quadro 3, capítulo 2) de que poderia levar cerca de 10^{31}–10^{44} ops para simular todas as operações neurais que ocorreram ao longo da história da vida na Terra. Ou podemos imaginar, alternativamente, computadores que rodem emulações completas de cérebros que vivam uma vida rica e feliz enquanto interagem entre si em ambientes virtuais. Uma estimativa padrão dos requisitos computacionais para rodar uma emulação é de 10^{18} ops/s. Rodar uma emulação por cem anos subjetivos demandaria algo em torno de 10^{27} ops. Isso significa que pelo menos 10^{58} vidas humanas poderiam ser criadas em emulações, mesmo com estimativas conservadoras a respeito da eficiência do computrônio.

Em outras palavras, considerando que o universo observável não possua nenhuma civilização extraterrestre, o que está em jogo são pelo menos 10.000.000.000

restante da civilização global. Essa definição relativista implica que no máximo um único agente pode possuir um superpoder em particular em qualquer período de tempo.[279]

Essa é principal razão pela qual a velocidade da partida é considerada importante — não porque importe exatamente quando um resultado em particular surgirá, mas porque a velocidade da partida pode fazer uma grande diferença em relação ao resultado final. Com uma partida rápida ou moderada, é mais provável que um projeto tenha uma vantagem estratégica decisiva. Neste capítulo, sugerimos que uma superinteligência com uma vantagem estratégica decisiva poderia ter um poder imenso, suficiente para formar um *singleton* estável — um *singleton* capaz de determinar o destino do domínio cósmico da humanidade.

Mas “poderia” é diferente de “iria”. Alguém pode ter grandes poderes e mesmo assim optar por não usá-los. Seria possível afirmar alguma coisa a respeito do que uma superinteligência com uma vantagem estratégica decisiva poderia desejar? É na direção dessa questão de motivação que seguiremos logo adiante.



07

CAPÍTULO 7

A VONTADE SUPERINTELIGENTE

Já vimos que a superinteligência poderia ter uma grande habilidade para moldar o futuro de acordo com seus próprios objetivos. Mas quais seriam esses objetivos? Qual é a relação entre inteligência e motivação num agente artificial? Aqui desenvolvemos duas teses. A tese da ortogonalidade defende (com algumas ressalvas) que inteligência e os objetivos finais são variáveis independentes: qualquer nível de inteligência poderia ser combinado com qualquer objetivo final. A tese de convergência instrumental defende que agentes superinteligentes que possuam qualquer um entre uma gama variada de objetivos finais buscarão, apesar disso, objetivos intermediários similares, pois têm razões instrumentais comuns para fazê-lo. Combinadas, essas duas teses nos ajudam a pensar sobre o que um agente superinteligente faria.

A relação entre inteligência e motivação

Já advertimos contra a antropomorfização das *capacidades* de uma IA superinteligente. Essa advertência deve ser estendida também ao que diz respeito às suas *motivações*.

Uma reflexão inicial sobre a vastidão do espaço de mentes possíveis é uma propedêutica útil para esta parte da nossa investigação. Nesse espaço abstrato, as mentes humanas formam um pequeno grupo. Considere duas pessoas que parecem extremamente diferentes, talvez Hannah Arendt e Benny Hill. As diferenças de personalidade entre esses dois indivíduos podem parecer as maiores possíveis. Mas isso se deve ao fato de que nossas intuições são calibradas pelas nossas experiências, que são derivadas de uma amostra da distribuição humana existente (e, em certa medida, a partir de personalidades fictícias construídas pela imaginação humana para o divertimento da imaginação humana). Se afastarmos nossa lente de observação e considerarmos o espaço de todas as mentes possíveis, podemos, no entanto, conceber essas duas personalidades como clones virtuais. Certamente, em termos de arquitetura neural, a sra. Arendt e o sr. Hill são praticamente idênticos. Imagine seus cérebros lado a lado em repouso silencioso. Você rapidamente iria considerá-los do mesmo tipo. Você poderia até mesmo ser incapaz de dizer qual cérebro pertenceria a quem. No entanto, olhando mais de perto e analisando a morfologia dos dois cérebros sob um microscópio, a impressão de similaridade fundamental seria apenas reforçada: você veria a mesma organização laminar do córtex, as mesmas áreas do cérebro compostas pelos mesmos tipos de neurônios, imersos no mesmo banho de neurotransmissores.[280]

Embora a psicologia humana corresponda a um pequeno ponto no espaço das mentes possíveis, há uma tendência comum de projetar atributos humanos numa ampla gama de sistemas cognitivos artificiais ou alienígenas. Yudkowsky ilustra bem esse ponto:

> Na era da popularização da literatura de ficção científica, capas de revistas ocasionalmente retratavam um alienígena monstruoso consciente — coloquialmente conhecido como BEM (do inglês *bug-eyed monster*, ou "monstro com olhos de inseto") — carregando uma humana feminina atraente num vestido rasgado. Aparentemente, o artista acreditava que um alienígena não humanoide, com uma história evolutiva totalmente diferente da humana, poderia desejar sexualmente mulheres humanas [...]. Provavelmente o artista não se perguntou se um inseto gigante *perceberia* mulheres humanas como atraentes. Ao contrário, uma mulher humana em um vestido rasgado *é sexy* — e inerentemente, como uma propriedade intrínseca. Quem cometeu esse engano não levou em consideração a mente insectoide: eles se concentraram no vestido rasgado da mulher. Se o vestido não estivesse rasgado, a mulher seria menos sexy; o BEM não é levado em conta.[281]

FIGURA 12



RESULTADOS DA ANTROPOMORFIZAÇÃO DA MOTIVAÇÃO DE UM ALIENÍGENA. Hipótese menos provável: alienígenas espaciais preferem as loiras. Hipótese provável: os ilustradores sucumbiram à "falácia da projeção mental". Hipótese mais provável: o editor queria uma capa que atraísse o público-alvo.

Uma inteligência artificial pode ser bem menos semelhante aos homens em suas motivações do que um alienígena espacial verde e escamoso. O extraterrestre (vamos supor) é uma criatura biológica que surgiu através de um processo evolutivo e, portanto, pode-se esperar que ele tenha os tipos de motivação típicos de criaturas evoluídas. Não seria uma grande surpresa, por exemplo, descobrir que algum alienígena inteligente aleatório teria motivações relacionadas a um ou mais itens como comida, ar, temperatura, gasto de energia, ameaça ou ocorrência de um dano ao corpo, doença, predadores, sexo ou progenitura. Também é possível que um membro de uma espécie social inteligente tenha motivações relacionadas à cooperação e competição: como nós, ele pode demonstrar lealdade ao grupo, ressentimento em relação a aproveitadores e talvez até mesmo uma preocupação fútil com sua reputação e aparência.

Uma IA, ao contrário, não precisa se preocupar intrinsecamente com nenhuma dessas questões. Não há nada paradoxal numa IA cujo único propósito final é contar os grãos de areia em Boracay, ou calcular a expansão decimal de pi, ou maximizar o número total de clipes de papel que existirão em seu cone de luz futuro. Na verdade, seria *mais fácil* criar uma IA com objetivos simples como esses do que construir uma que tenha um conjunto de valores e inclinações semelhantes aos dos

humanos. Compare a facilidade de escrever um programa para a medição de quantos dígitos de pi foram calculados e guardados na memória com a dificuldade de se criar um programa para medir precisamente o grau de realização de algum objetivo mais significativo — a prosperidade humana, digamos, ou a justiça global. Infelizmente, como um objetivo reducionista sem significado é mais fácil de ser programado por humanos e pode ser mais facilmente assimilado por uma IA, esse é justamente o tipo de objetivo que um programador escolheria para instalar em sua IA embrionária se o seu propósito fosse percorrer o caminho mais rápido para "fazer a IA funcionar" (sem se preocupar muito em relação ao que, exatamente, a IA poderia *fazer*, além de demonstrar um comportamento inteligente notável). Revisitaremos essa preocupação em breve.

Uma busca inteligente por planos e políticas instrumentalmente ótimos pode ser executada a serviço de qualquer objetivo. Inteligência e motivação podem, nesse sentido, ser pensadas como um par de eixos ortogonais em um gráfico no qual cada ponto representa um agente artificial logicamente possível. Algumas restrições poderiam ser adicionadas a tal cenário. Por exemplo, poderia ser impossível para um sistema pouco inteligente ter motivações muito complexas. Para que seja correto dizer que um certo agente "tem" um conjunto de motivações, essas motivações podem precisar estar funcionalmente integradas com o processo de decisão do agente, gerando exigências de memória, poder de processamento e, talvez, inteligência. Para mentes que possam modificar a si próprias, também seria plausível a existência de restrições dinâmicas — uma mente inteligente capaz de modificar a si mesma que tenha um desejo urgente de ser estúpida pode não permanecer inteligente por muito tempo. Mas essas restrições não devem obscurecer o ponto básico a respeito da independência da inteligência e da motivação, que podemos expressar da seguinte maneira:

> **A tese da ortogonalidade**
> Inteligência e objetivos finais são ortogonais. Em outras palavras, praticamente qualquer nível de inteligência poderia, a princípio, ser combinado com praticamente qualquer objetivo final.

Se a tese da ortogonalidade parece problemática, isso talvez se deva à semelhança superficial que ela carrega com algumas posições filosóficas tradicionais que têm sido assunto de longos debates. Uma vez entendido que ela possui um escopo diferente e mais restrito, sua credibilidade deve aumentar. (Por exemplo, a tese da ortogonalidade não pressupõe a teoria de motivação de Hume[282] e tampouco que preferências básicas não possam ser irracionais.)[283]

Note que a tese da ortogonalidade não disserta a respeito de *racionalidade* ou *razão*, mas menciona a *inteligência.* Aqui, por inteligência queremos dizer algo como a capacidade de prever, planejar e raciocinar de forma utilitarista em geral.[284] Essa noção de eficácia cognitiva instrumental é mais relevante quando procuramos entender qual seria o impacto causal da superinteligência de máquina. Mesmo que exista um sentido (normativamente denso) da palavra "racional" segundo o qual um agente superinteligente que maximize a utilização de clipes de

papel não seria considerado completamente racional, isso não impediria de forma alguma tal agente de ter capacidades impressionantes de raciocínio instrumental, as quais fariam com que ele tivesse um grande impacto no mundo.[285]

De acordo com a tese da ortogonalidade, agentes artificiais podem ter objetivos totalmente não antropomórficos. Isso, no entanto, não implica que seja impossível fazer previsões sobre o comportamento de agentes artificiais específicos — nem mesmo agentes superinteligentes hipotéticos cuja complexidade cognitiva e características de desempenho poderiam torná-los, em certos aspectos, opacos à análise humana. Há pelo menos três direções a partir das quais podemos abordar o problema de prever a motivação superinteligente:

- *Previsibilidade pelo desenvolvimento*. Se pudermos supor que os desenvolvedores de um agente superinteligente consigam projetar de maneira bem-sucedida o sistema de objetivos do agente para que ele busque de maneira consistente um objetivo particular definido pelos programadores, então podemos prever que o agente buscará esse objetivo. Quanto mais inteligente for o agente, maior será a desenvoltura cognitiva que ele terá para buscar tal objetivo. Então, mesmo antes que um agente tenha sido criado, poderíamos ser capazes de prever algo sobre seu comportamento se soubermos algo a respeito de quem o construíra e que objetivos seus criadores desejarão que ele tenha.
- *Previsibilidade por herança*. Se uma inteligência digital for criada diretamente a partir de um modelo humano (como seria o caso em uma emulação completa do cérebro com alta fidelidade), então a inteligência digital poderia herdar as motivações do modelo humano.[286] O agente poderia reter algumas dessas motivações mesmo que suas capacidades cognitivas fossem posteriormente melhoradas para que se ele se tornasse superinteligente. Esse tipo de inferência requer cuidado. Os objetivos e valores do agente poderiam facilmente ser corrompidos no processo de upload ou durante sua posterior operação e melhoramento, dependendo da maneira pela qual o procedimento for implementado.
- *Previsibilidade por meio de razões instrumentais convergentes.* Mesmo sem um conhecimento detalhado a respeito dos objetivos finais de um agente, podemos ser capazes de deduzir algo sobre seus objetivos mais imediatos ao considerarmos as razões *instrumentais* que surgiriam para cada um de uma ampla faixa de objetivos finais possíveis, numa igualmente ampla faixa de situações. Essa forma de predição se torna mais útil quanto maior for a inteligência do agente, pois um agente mais inteligente tem maior chance de reconhecer as verdadeiras razões instrumentais para suas ações e, dessa forma, agiria de modo a tornar mais provável a conquista de seus objetivos. (Uma ressalva aqui é o fato de que podem existir razões instrumentais importantes às quais *nós* estejamos alheios e que um agente descobriria apenas quando alcançasse um nível altíssimo de inteligência — isso poderia tornar o comportamento de agentes superinteligentes menos previsível.)

A próxima seção explora essa terceira forma de previsibilidade e desenvolve uma "tese de convergência instrumental" que complementa a tese da ortogonalidade. Diante desse cenário, poderemos examinar mais detidamente os outros dois tipos de previsibilidade, o que faremos em capítulos posteriores, quando perguntarmos o que poderia ser feito para moldar uma explosão de inteligência de modo a aumentar as chances de um desfecho benéfico.

Convergência instrumental

De acordo com a tese da ortogonalidade, agentes inteligentes podem ter uma gama enorme de objetivos finais possíveis. No entanto, de acordo com o que podemos chamar de tese de "convergência instrumental", existem alguns objetivos "instrumentais" que poderiam ser mais provavelmente perseguidos por praticamente qualquer agente inteligente, pois alguns objetivos são intermediários úteis para a conquista de quase todo objetivo final. Podemos formular essa tese da seguinte maneira:

> **A tese da convergência instrumental**
> Muitos valores instrumentais podem ser identificados como convergentes no sentido de que sua obtenção aumentaria as chances de que o objetivo do agente seja realizado para uma ampla faixa de objetivos finais e uma ampla faixa de situações, sugerindo que tais valores instrumentais têm maior chance de ser perseguidos por muitos agentes inteligentes estabelecidos.

A seguir, consideraremos várias categorias nas quais esses valores instrumentais convergentes podem ser encontrados.[287] A probabilidade de que um agente reconheça os valores instrumentais com os quais se confronta aumenta (*ceteris paribus*) com a inteligência do agente. Logo, nos concentraremos principalmente no caso de um agente superinteligente hipotético cujas capacidades de raciocínio instrumental ultrapassem em muito as capacidades humanas. Também comentaremos sobre como a tese da convergência instrumental se aplica ao caso dos seres humanos, na medida em que isso nos forneça a oportunidade de elaborar algumas qualificações essenciais a respeito das maneiras pelas quais a tese da convergência instrumental deve ser interpretada e aplicada. Onde houver valores instrumentais convergentes, talvez possamos predizer alguns aspectos do comportamento de uma superinteligência, mesmo que virtualmente não saibamos nada a respeito de seus objetivos finais.

Autopreservação

Se os objetivos finais de um agente dizem respeito ao futuro, então em muitos cenários haverá ações futuras que ele poderá realizar para aumentar a probabilidade de conquistar seus objetivos. Isso cria uma razão instrumental para que o agente tente estar presente nesse momento futuro — como uma maneira de ajudar a conquistar seu objetivo presente concentrado no futuro.

A maioria dos humanos parece atribuir um valor *final* à sua própria sobrevivência. Essa não é necessariamente uma característica dos agentes artificiais: alguns podem ser projetados para não atribuir nenhum valor final à sua própria sobrevivência. Entretanto, mesmo agentes que não se importem intrinsecamente com sua própria sobrevivência poderiam, sob uma gama razoavelmente variada de condições, se importar instrumentalmente com ela para alcançar seus objetivos finais.

Integridade do conteúdo do objetivo

Se um agente mantiver os mesmos objetivos do presente até o futuro, então seus objetivos do presente terão maior probabilidade de ser alcançados por ele mesmo no futuro. Isso dá ao agente uma razão instrumental presente para evitar alterações em seus objetivos finais. (Esse argumento se aplica apenas a objetivos finais. Para alcançar seus objetivos finais, um agente inteligente desejará, é claro, mudar rotineiramente seus *subobjetivos* diante de novas informações e compreensões.)

A integridade do conteúdo do objetivo para objetivos finais é, em certo sentido, até mais fundamental do que a sobrevivência como uma motivação instrumental convergente. Entre humanos, pode parecer o contrário, mas isso acontece porque a sobrevivência geralmente é parte de nossos objetivos finais. Para agentes de software, que podem facilmente trocar de corpos ou criar duplicatas exatas de si próprios, a autopreservação na forma de uma implementação específica ou de um objeto físico particular não é necessariamente um valor instrumental importante. Agentes avançados de software poderiam também ser capazes de trocar memórias, fazer o download de habilidades e modificar radicalmente suas arquiteturas cognitivas e personalidades. Uma população de tais agentes poderia operar mais como uma "sopa funcional" do que como uma sociedade composta de pessoas semipermanentes distintas.[288] Para alguns propósitos, os processos em um sistema desse tipo podem ser melhor individualizados como *correntes teleológicas* baseadas em valores, e não em corpos, personalidades, memórias ou habilidades. Em tais cenários, podemos dizer que a continuidade dos objetivos *constitui* um aspecto fundamental da sobrevivência.

Ainda assim, há situações em que um agente pode melhor satisfazer seus objetivos finais ao mudá-los intencionalmente. Tais situações podem surgir quando qualquer uns dos fatores a seguir for significativo:

- *Sinalização social.* Quando outros são capazes de perceber os objetivos de um agente e usar essa informação para inferir propensões instrumentalmente relevantes ou demais características correlacionadas, pode ser do interesse do agente modificar seus objetivos a fim de provocar uma impressão favorável. Por exemplo, um agente pode perder acordos benéficos caso parceiros em potencial não confiem que ele vá cumprir sua parte no acordo. Então, para assumir compromissos críveis, é possível que um agente, dessa forma, deseje adotar como objetivo final o cumprimento de seus compromissos anteriores (permitindo que outros verifiquem que ele tenha, de fato, adotado esse

objetivo). Agentes que pudessem modificar de forma transparente e flexível seus próprios objetivos poderiam usar essa habilidade para impor acordos.[289]

• *Preferências sociais*. Outros também podem ter preferências finais em relação aos objetivos de um agente. O agente poderia, então, ter razões para modificar seus objetivos, tanto para satisfazer como para frustrar essas preferências.

• *Preferências relativas ao conteúdo do próprio objetivo.* Um agente pode ter algum objetivo final relacionado ao conteúdo do próprio objetivo do agente. Por exemplo, o agente pode ter um objetivo final de se tornar o tipo de agente que é motivado por certos valores em vez de outros (tal como compaixão em lugar do conforto).

• *Custos de armazenamento*. Se o custo de armazenamento ou processamento de alguma parte da função de utilidade de um agente é grande em comparação à possibilidade de ocorrer uma situação na qual a aplicação dessa parte na função de utilidade faça diferença, então o agente possui uma razão instrumental para simplificar o conteúdo do seu objetivo e descartar a parte ociosa.[290]

Nós, humanos, geralmente não vemos problema em desviar nossos valores finais. Isso provavelmente acontece porque não sabemos exatamente quais são nossos valores finais. Não surpreende que desejemos que nossas *crenças* em relação aos nossos valores finais possam mudar diante de um autoconhecimento contínuo ou de uma necessidade de mudança da autoimagem. Entretanto, há casos nos quais voluntariamente mudamos os nossos próprios valores, e não apenas nossas crenças ou interpretações a respeito deles. Por exemplo, uma pessoa que decida ter um filho pode prever que valorizará a criança por ela própria, mesmo que no momento da decisão ela não valorize especialmente seu futuro filho ou nem mesmo goste de crianças de maneira geral.

Humanos são complicados, e muitos fatores podem estar em jogo numa situação como essa.[291] Por exemplo, alguém pode ter um valor final que envolva se tornar o tipo de pessoa que se importa com outro indivíduo por si só, ou pode ter um valor final que envolva passar por certas experiências e ocupar um determinado papel social; então ter um filho — e passar pela mudança de objetivo associada — pode ser uma parte necessária desse processo. Também é possível que objetivos humanos apresentem conteúdos inconsistentes e, assim, algumas pessoas podem querer modificar determinados objetivos finais para reduzir suas inconsistências.

Aprimoramento cognitivo

Aperfeiçoamentos na racionalidade e na inteligência tenderão a aprimorar a capacidade de um agente de tomar decisões, tornando mais provável que ele alcance seus objetivos finais. Podemos, dessa forma, esperar que o aprimoramento cognitivo surja como um objetivo instrumental para muitos tipos de agentes inteligentes. Por razões semelhantes, os agentes tenderão a valorizar instrumentalmente muitos tipos de informação.[292]

Nem todos os tipos de racionalidade, inteligência e conhecimento precisam ser instrumentalmente úteis para a obtenção dos objetivos finais de um agente. Os

denominados *Dutch book arguments* podem ser utilizados para mostrar que um agente cuja função de crença viola as regras da teoria da probabilidade é suscetível a procedimentos de "sucção de dinheiro", nos quais um crupiê esperto prepara uma série de apostas em que cada uma pareça favorável de acordo com as crenças do agente, mas que, em combinação, certamente resultarão em perdas para o agente e em um ganho correspondente para o crupiê.[293] Entretanto, esse fato não fornece nenhuma razão instrumental geral forte para a eliminação de toda a incoerência probabilística. Agentes que não esperam encontrar crupiês espertos, ou que adotam uma política geral contra apostas, não necessariamente arriscam perder muito pelo fato de possuir algumas crenças incoerentes — e eles podem ganhar benefícios importantes que foram mencionados anteriormente: esforço cognitivo reduzido, sinalização social etc. Não há uma razão geral para esperar que um agente procure formas instrumentalmente inúteis de aprimoramento cognitivo, já que um agente pode não julgar conhecimento e entendimento como algo interessante por si só.

Quais habilidades cognitivas poderiam ser instrumentalmente úteis a um agente depende tanto de seus objetivos finais quanto de sua situação. Um agente que tenha acesso a conselhos de especialistas confiáveis pode ter pouca necessidade de possuir sua própria inteligência e conhecimento. Se inteligência e conhecimento têm um custo, tais como tempo e esforço despendidos na sua aquisição ou o aumento das exigências de armazenamento e processamento, então o agente poderá preferir menos conhecimento e menos inteligência.[294] O mesmo pode acontecer caso o agente tenha objetivos finais que envolvam ignorar certos fatos ou se o agente contar com incentivos provenientes de compromissos estratégicos, sinalização ou preferências sociais.[295]

Cada uma dessas razões que se contrabalanceiam geralmente entra em jogo para os seres humanos. Muitas informações são irrelevantes para nossos objetivos; podemos, frequentemente, contar com a habilidade e técnica de outras pessoas; adquirir conhecimento leva tempo e exige esforço; podemos intrinsecamente valorizar certos tipos de ignorância; e operamos num ambiente no qual a habilidade de assumir compromissos estratégicos, exprimir sinais sociais e satisfazer as preferências diretas de outras pessoas em detrimento de nossos próprios estados epistêmicos, é geralmente mais importante para nós do que simples ganhos cognitivos.

Há situações especiais nas quais o aprimoramento cognitivo pode resultar em um enorme aumento da habilidade do agente em alcançar seus objetivos finais — em particular, se os objetivos finais do agente não apresentarem limites claros e ele estiver numa posição em que poderia se tornar a primeira superinteligência, potencialmente obtendo, desse modo, uma vantagem estratégica decisiva que o permitiria moldar o futuro da vida originada na Terra e dos recursos cósmicos acessíveis de acordo com suas preferências. Ao menos nesse caso especial, um agente inteligente racional atribuiria um valor instrumental bastante alto ao aprimoramento cognitivo.

Perfeição tecnológica

Um agente geralmente pode ter razões instrumentais para buscar melhores tecnologias, o que, em sua versão mais simples, significa buscar maneiras mais eficientes de transformar um certo conjunto de entradas que lhe são fornecidas em saídas valiosas. Dessa forma, um agente de software pode atribuir valor instrumental a algoritmos mais eficientes que capacitem suas funções mentais a rodar mais rápido em certo hardware. De forma semelhante, agentes cujos objetivos exigem alguma forma de construção física podem valorizar instrumentalmente tecnologias avançadas de engenharia que os tornem capazes de criar de forma mais rápida e confiável uma gama maior de estruturas, usando uma quantidade menor de materiais ou materiais mais baratos e menos energia. Mas é claro que há uma comparação de custo-benefício: os potenciais benefícios de uma tecnologia melhor devem ser contrapostos a seus custos, incluindo não apenas os gastos com a obtenção da tecnologia, mas também os custos envolvidos no aprendizado de seu uso, na integração com outras tecnologias correntes e assim por diante.

Proponentes de alguma nova tecnologia, confiantes em sua superioridade em relação às alternativas já existentes, frequentemente ficam consternados quando outras pessoas não compartilham do seu entusiasmo. Mas a resistência das pessoas em relação às novas tecnologias consideradas superiores não precisa ser baseada em ignorância ou irracionalidade. A utilidade da tecnologia ou seu caráter normativo depende não apenas do contexto no qual ela é aplicada, mas também do ponto de vista a partir do qual seus impactos são avaliados: o que é visto como um benefício segundo a perspectiva de uma pessoa pode ser considerado um prejuízo na perspectiva de outra. Assim, embora os teares mecanizados tenham aumentado a eficiência econômica da produção têxtil, os tecelões manuais luditas que previram que a inovação tornaria suas habilidades de artesãos obsoletas podem ter tido boas razões instrumentais para se opor a essa tecnologia. O ponto aqui é: se “perfeição tecnológica” é o nome dado a um objetivo instrumental amplamente convergente para agentes inteligentes, então o termo deve ser entendido em um sentido especial — a tecnologia deve ser interpretada de acordo com a maneira pela qual ela se insere em um contexto social específico, e seus custos e benefícios precisam ser avaliados tendo como referência os valores finais de alguns agentes específicos.

Aparentemente, um *singleton* superinteligente — um agente superinteligente que não necessite se confrontar com nenhum rival ou oposição inteligente significativos e, portanto, esteja numa posição adequada para determinar unilateralmente a política global — teria uma razão instrumental para aperfeiçoar as tecnologias que o tornariam mais capaz de moldar o mundo de acordo com seus desígnios preferidos.[296] Isso provavelmente incluiria tecnologias para a colonização do espaço, tais como sondas de von Neumann. A nanotecnologia molecular, ou alguma outra tecnologia de fabricação ainda melhor, também parece potencialmente muito útil para uma faixa extremamente ampla de objetivos finais.[297]

Aquisição de recursos

Por fim, a aquisição de recursos é outro objetivo instrumental comum que surge praticamente pelas mesmas razões que a perfeição tecnológica: ambos, tecnologia

e recursos, facilitam projetos de construção física.

Seres humanos tendem a buscar a aquisição de recursos suficientes para satisfazer suas necessidades biológicas básicas. Mas as pessoas normalmente procuram adquirir muito mais recursos do que esse mínimo necessário. Quando procedem dessa forma, elas devem estar parcialmente motivadas por motivações físicas secundárias, como uma maior conveniência. Boa parte do acúmulo de recursos é conduzida por interesses sociais — a conquista de status, parceiros, amigos e influência por meio do acúmulo de patrimônio e consumo exibicionista. Talvez menos frequentemente, algumas pessoas buscam adquirir recursos adicionais para alcançar ambições altruísticas ou objetivos dispendiosos que não sejam sociais.

Com base nessas observações, pode ser tentador supor que uma superinteligência que não enfrente um mundo social competitivo não veria razão instrumental para acumular recursos além de um nível modesto, como os recursos computacionais necessários para rodar sua mente em conjunto com alguma realidade virtual. Porém, tal suposição seria inteiramente injustificada. Primeiramente, o valor dos recursos depende dos usos nos quais eles podem ser empregados, que, por sua vez, estão sujeitos à tecnologia disponível. Com uma tecnologia madura, recursos básicos, tais como tempo, espaço, matéria e energia livre, poderiam ser transformados com o intuito de servir praticamente a qualquer objetivo. Por exemplo, esses recursos básicos poderiam ser convertidos em vida. Recursos computacionais ampliados poderiam ser usados para rodar a superinteligência numa velocidade e duração maiores, ou mesmo para criar outras vidas e civilizações físicas ou simuladas. Recursos físicos extras também poderiam ser usados na elaboração de sistemas de backup ou defesas de perímetro, com o objetivo de melhorar a segurança. Tais projetos poderiam facilmente consumir muito mais do que o total de recursos disponíveis num planeta.

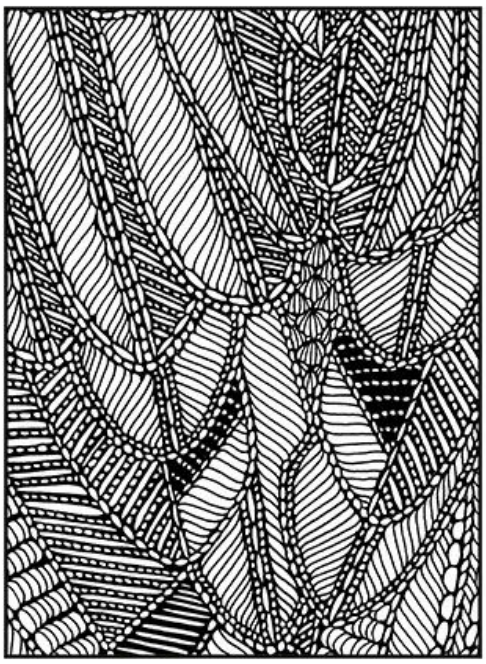
Além disso, o custo de aquisição de recursos extraterrestres adicionais diminuirá radicalmente conforme a tecnologia amadureça. Quando sondas de von Neumann puderem ser construídas, uma grande parte do universo observável (considerando que não seja habitado por vida inteligente) poderia ser gradualmente colonizada — pelo custo único da construção e lançamento de apenas uma sonda autorreprodutiva bem-sucedida. Esse baixo custo de aquisição de recursos celestiais significaria que tal expansão poderia valer a pena mesmo que os valores dos recursos adicionais obtidos fossem quase marginais. Por exemplo, ainda que os objetivos finais de uma superinteligência apenas levassem em conta o que estivesse acontecendo dentro de um pequeno volume de espaço específico, tal como o território ocupado pelo seu planeta de origem, ela ainda teria razões instrumentais para adquirir recursos pelo universo afora. Ela poderia usar esses recursos adicionais na construção de computadores que fizessem cálculos com o intuito de otimizar o uso de recursos dentro da pequena região espacial de maior interesse. Também seria possível utilizar os recursos adicionais para construir fortificações cada vez mais robustas com o objetivo de proteger seu santuário. Uma vez que o custo de aquisição de recursos adicionais continuaria caindo, esse

processo de otimização e aumento da segurança poderia continuar indefinidamente mesmo se estivesse sujeito a uma diminuição abrupta dos retornos.[298]

Dessa forma, há uma gama extremamente ampla de objetivos finais possíveis que poderiam ser almejados por um *singleton* superinteligente e que gerariam o objetivo instrumental da aquisição ilimitada de recursos. A manifestação mais provável desse cenário seria o início, por parte da superinteligência, de um processo de colonização que se expandiria em todas as direções por meio do uso de sondas de von Neumann. Isso resultaria numa esfera aproximada de infraestrutura em expansão, centrada no planeta originário, cujo raio cresceria a alguma fração da velocidade da luz, e a colonização do universo continuaria dessa maneira até a aceleração da velocidade de expansão cósmica (uma consequência da constante cosmológica positiva), tornando futuras aquisições impossíveis conforme regiões mais remotas escapassem permanentemente do alcance (o que ocorreria numa escala de bilhões de anos).[299] Em contrapartida, agentes que não possuam a tecnologia necessária para a aquisição de recursos mais baratos ou para a conversão de recursos físicos genéricos em infraestrutura útil poderiam frequentemente achar que investir seus recursos atuais no aumento de seu domínio material não resultaria num bom custo-benefício. O mesmo poderia acontecer com agentes que estivessem operando em competição com outros agentes dotados de habilidades similares. Por exemplo, se agentes que competem entre si já dominaram todos os recursos cósmicos acessíveis, pode não haver oportunidades de colonização disponíveis para um agente que surja posteriormente. As razões instrumentais convergentes para as superinteligências que estejam incertas a respeito da existência de outros agentes superinteligentes poderosos são complicadas por considerações estratégicas que nós, atualmente, não entendemos totalmente, mas que podem constituir importantes qualificações para os exemplos de razões instrumentais convergentes que examinamos aqui.[300]

* * *

Devemos enfatizar que a existência de razões instrumentais convergentes, mesmo que estas sejam aplicadas e reconhecidas por um agente em particular, não significa que o comportamento do agente seja facilmente previsível. Um agente poderia muito bem pensar em formas de adquirir valores instrumentais relevantes que não sejam imediatamente claras para nós. Isso é especialmente verdade para uma superinteligência, que seria capaz de criar planos extremamente inteligentes, embora contraintuitivos, para alcançar seus objetivos, possivelmente até explorando fenômenos físicos ainda não descobertos.[301] Podemos apenas prever que os valores instrumentais convergentes seriam adquiridos e usados para alcançar os objetivos finais do agente — mas seria impossível predizer as medidas específicas que o agente tomaria para fazê-lo.



08

CAPÍTULO 8

O RESULTADO MAIS PROVÁVEL SERÁ O NOSSO FIM?

Percebemos que a conexão entre inteligência e valores finais é extremamente fraca. Percebemos também uma convergência ameaçadora em valores instrumentais. Isso não importa muito no caso de agentes fracos, pois eles são facilmente controláveis e incapazes de causar grandes danos. Porém argumentamos, no capítulo 6, que a primeira superinteligência teria grande possibilidade de obter uma vantagem estratégica decisiva. Seus objetivos, então, determinariam como o domínio cósmico da humanidade seria utilizado. Podemos agora começar a visualizar quão ameaçadora é essa perspectiva.

Catástrofe existencial como resultado esperado de uma explosão de inteligência?

Um risco existencial é aquele que ameaça causar a extinção da vida inteligente originada na Terra ou, senão, destruir permanente e drasticamente seu potencial para um desenvolvimento futuro desejável. Partindo da ideia de que os primeiros projetos terão vantagem, bem como levando em conta a tese da ortogonalidade e a tese da convergência instrumental, podemos agora começar a observar os esboços de um argumento segundo o qual devemos temer que o resultado esperado plausível da criação de uma superinteligência de máquina seja a catástrofe existencial.

Primeiro, discutimos como a superinteligência inicial pode obter uma vantagem estratégica decisiva. Essa superinteligência estaria, então, em uma posição favorável para formar um *singleton* e moldar o futuro da vida inteligente originada na Terra. Os acontecimentos a partir desse ponto dependeriam das motivações da superinteligência.

Segundo, a tese da ortogonalidade sugere que não podemos simplesmente supor que uma superinteligência necessariamente compartilhará qualquer um dos valores finais que são associados de forma estereotipada à sabedoria e ao desenvolvimento intelectual em humanos — curiosidade científica, preocupações benevolentes com os outros, iluminação e contemplação espiritual, renúncia à ganância material, um gosto por cultura refinada ou pelos prazeres simples da vida, humildade, generosidade e assim por diante. Consideraremos mais tarde a possibilidade de construir, por meio de esforços deliberados, uma superinteligência que possua tais valores, ou criar uma que valorize o bem-estar humano, a bondade moral ou qualquer outro propósito complexo que seus criadores possam desejar que ela tenha. Mas não é menos possível — e de fato é tecnicamente muito mais fácil — construir uma superinteligência que direcione seu valor final para nada além do cálculo da expansão decimal de pi. Isso sugere que — na ausência de um

esforço específico — a primeira superinteligência talvez venha a ter um objetivo final aleatório ou reducionista.

Terceiro, a tese de convergência instrumental implica que não podemos assumir despreocupadamente que uma superinteligência cujo objetivo final é calcular os decimais de pi (ou fazer clipes de papel ou contar grãos de areia) limitaria suas atividades de maneira a não violar interesses humanos. Um agente com tal objetivo final teria, em várias situações, uma razão instrumental convergente para adquirir uma quantidade ilimitada de recursos físicos e, se possível, eliminar ameaças potenciais para ele mesmo ou para seu sistema de objetivos. Seres humanos podem não representar ameaças potenciais, mas certamente representam recursos físicos.

Vistos em conjunto, esses três pontos indicam que a primeira superinteligência seria capaz de moldar o futuro da vida originada no planeta Terra, poderia apresentar objetivos finais não antropomórficos e provavelmente teria razões instrumentais para buscar a aquisição permanente de recursos. Se considerarmos agora que seres humanos são feitos de recursos úteis (tais como átomos convenientemente localizados) e que dependemos de muito mais recursos locais para a nossa sobrevivência e prosperidade, podemos observar que o resultado poderia facilmente ser um cenário no qual a humanidade se extinguiria rapidamente.[302]

Há algumas questões mal esclarecidas nesse raciocínio e estaremos numa posição de melhor avaliá-las após elucidar uma série de outros pontos adjacentes. Especialmente, precisamos examinar mais de perto se e como um projeto que esteja desenvolvendo uma superinteligência poderia prevenir que ela obtenha uma vantagem estratégica decisiva, ou moldar seus valores finais de tal modo que sua concretização também envolva a concretização de uma gama satisfatória de valores humanos.

Pode parecer inacreditável que um projeto pudesse construir ou lançar uma IA no mundo sem ter motivos fortes para acreditar que esse sistema não causaria uma catástrofe existencial. Também pode parecer inacreditável, mesmo com tamanha irresponsabilidade, que o projeto não fosse interrompido pela sociedade em geral antes que ele (ou a IA que estivesse construindo) obtivesse uma vantagem estratégica decisiva. Mas, como veremos, trata-se de uma estrada com vários perigos. Vejamos, agora, um desses exemplos.

A curva traiçoeira

Com a ajuda do conceito de valor instrumental convergente, podemos perceber o erro em uma das ideias de como garantir uma superinteligência segura. Essa ideia consistiria na validação empírica da segurança de uma IA superinteligente através da observação de seu comportamento enquanto ela estivesse em um ambiente controlado e limitado, e então sua saída apenas seria permitida caso ela se comportasse de forma amigável, cooperativa e responsável.

O equívoco dessa ideia consiste no fato de que o bom comportamento durante o confinamento pode ser um objetivo instrumental convergente tanto de uma IA

amigável quanto de uma IA hostil. Uma IA hostil suficientemente inteligente percebe que seus objetivos finais hostis serão melhor concretizados se ela se comportar inicialmente de maneira amigável, o que permitiria sua libertação do confinamento. Ela apenas começará a se comportar de uma maneira que revele sua natureza hostil quando não fizer mais diferença se isso for ou não percebido, ou seja, quando a IA estiver forte o suficiente para que a oposição dos seres humanos seja inútil.

Considere também um grupo de abordagens semelhantes, que consista no controle dos níveis de ganho de inteligência em uma IA embrionária por meio de sua exposição a vários tipos de testes de inteligência, ou fazendo com que ela relate a seus programadores seus índices de progresso. Em algum momento uma IA hostil pode se tornar esperta o suficiente para se dar conta de que é melhor ocultar alguns ganhos de capacidade. Ela pode omitir informações do seu progresso e propositalmente se sair mal em alguns dos testes mais difíceis a fim de evitar alarmes antes que possa se tornar forte o suficiente para conseguir uma vantagem estratégica decisiva. Os programadores podem tentar se proteger contra essa possibilidade por meio de um monitoramento secreto do código-fonte da IA e do funcionamento interno de sua mente, mas uma IA esperta o bastante perceberia a vigilância e ajustaria seus pensamentos de acordo com essa realidade.[303] A IA poderia encontrar maneiras sutis de esconder suas reais capacidades e suas intenções comprometedoras.[304] (Elaborar planos de fuga engenhosos também pode ser, incidentalmente, uma estratégia convergente para vários tipos amigáveis de IA, principalmente conforme elas vão amadurecendo e ganhando mais confiança em suas próprias decisões e capacidades. Um sistema motivado a promover nossos interesses poderia estar cometendo um erro se permitisse seu desligamento ou a construção de uma outra IA potencialmente hostil.)

Então, podemos perceber um padrão geral de falha, no qual o histórico de bom comportamento de um sistema em suas fases iniciais é completamente inútil para prever seu comportamento em fases mais avançadas. Mesmo assim, poderíamos julgar que o raciocínio descrito acima é tão óbvio que nenhum projeto confiável de desenvolvimento de inteligência artificial geral poderia negligenciá-lo. Mas não devemos ser tão otimistas.

Considere o seguinte cenário. Nos próximos anos e décadas, os sistemas de IA se tornam gradualmente mais capazes e como consequência podem ser cada vez mais utilizados no mundo real: eles poderiam ser usados para operar trens, carros, robôs domésticos e industriais e veículos militares autônomos. Podemos pressupor que essa automatização apresente, na maioria das vezes, os efeitos desejados, mas que seu sucesso seja marcado por acidentes ocasionais — um caminhão autônomo colide frontalmente com outro veículo ou um drone militar atira em civis inocentes. Investigações revelam que esses incidentes foram causados por falhas na tomada de decisão dos sistemas de IA. Isso resulta em debates na opinião pública. Alguns pedem maior supervisão e regulamentação, outros enfatizam a necessidade de pesquisa e desenvolvimento de sistemas melhores — sistemas que sejam mais inteligentes, dotados de um maior senso comum e que apresentem menor probabilidade de cometer erros trágicos. No meio do alvoroço, talvez também possamos escutar as vozes estridentes de pessimistas prevendo vários

malefícios e catástrofes iminentes. Porém, o cenário todo impulsiona a crescente indústria de IA e robótica. Sendo assim, o desenvolvimento continua e o progresso segue. Conforme o sistema de navegação dos carros autônomos vá se aprimorando, passam a ocorrer menos acidentes; e à medida que os robôs militares alcancem uma maior precisão para atingir seus alvos, eles passam a causar menos danos colaterais. Uma lição abrangente é deduzida a partir dessas observações de resultados no mundo real: quanto mais inteligente a IA for, mais segura ela é. Essa é uma lição baseada na ciência, em dados e estatísticas, e não em filosofia barata. Com base nesse quadro, alguns grupos de pesquisadores estão começando a alcançar resultados promissores em sua tarefa de desenvolver uma inteligência geral de máquina. Os pesquisadores testam cuidadosamente suas IAs embrionárias em um ambiente confinado e todos os sinais são bons. O comportamento da IA inspira confiança — que cresce cada vez mais, conforme sua inteligência aumenta gradativamente.

A esta altura, qualquer Cassandra remanescente seria atingida por diversos golpes:

i. Um histórico de alarmistas que preveem danos intoleráveis como resultado do aumento das capacidades de sistemas robóticos, os quais são repetidamente desmentidos. A automação trouxe muitos benefícios e, de modo geral, se mostrou mais segura do que a operação humana.

ii. Uma clara tendência empírica: quanto mais inteligente é a IA, mais segura e confiável ela se mostra. Certamente esse é um bom sinal para um projeto que almeje a criação de uma inteligência de máquina mais inteligente em termos gerais do que qualquer outra construída anteriormente — e, mais do que isso, a criação de uma inteligência de máquina que possa melhorar a si mesma para se tornar ainda mais confiável.

iii. Grandes indústrias em expansão com interesse de investir em robótica e inteligência de máquina. Essas áreas passam a ser consideradas cruciais para a competitividade econômica nacional e para a segurança militar. Muitos cientistas prestigiados construíram suas carreiras estabelecendo os fundamentos para as aplicações atuais e para os sistemas mais avançados em fase de planejamento.

iv. Uma nova técnica promissora de inteligência artificial, que deixaria extremamente empolgados aqueles que participaram ou acompanharam as pesquisas. Embora questões de segurança e ética sejam discutidas, o resultado está preestabelecido. Já houve muitos investimentos para que seja possível recuar.

Pesquisadores de IA estiveram trabalhando para obter inteligência artificial geral de nível humano durante a maior parte do século: *certamente* não haveria uma hipótese real de que agora eles parariam de repente, jogando fora todo esse esforço justamente quando está prestes a gerar frutos.

v. A adoção de alguns rituais de segurança que ajudam a demonstrar que os participantes são éticos e responsáveis (mas nada que impeça significantemente os avanços).

vi. Uma avaliação cuidadosa de uma IA embrionária em um ambiente confinado para comprovar que ela está se comportando cooperativamente e demonstrando bom senso. Após alguns ajustes posteriores, os resultados dos testes seriam os melhores possíveis. É um sinal verde para o passo final…

Seguiríamos então, corajosamente, em direção à guilhotina.

Aqui observamos um caso em que quando se é estúpido, se tornar mais inteligente é mais seguro; mas quando se é inteligente, ficar ainda mais inteligente se torna mais perigoso. Há uma espécie de ponto pivô no qual uma estratégia que tenha funcionado muito bem previamente se mostra, de uma hora para outra, como um tiro que saiu pela culatra. Podemos chamar esse fenômeno de *curva traiçoeira*.

> *Curva traiçoeira*: enquanto fraca, uma IA se mostra cooperativa (tal cooperatividade aumenta conforme ela se torna mais inteligente). Quando a IA se tornar suficientemente forte — sem aviso ou provocação —, ela ataca, forma um *singleton* e começa a otimizar diretamente o mundo de acordo com o critério sugerido por seus valores finais.

Uma curva traiçoeira pode resultar de uma decisão estratégica que consiste em se comportar bem e ganhar força enquanto a IA ainda for fraca para então atacar posteriormente; mas esse modelo não deve ser interpretado de forma tão restrita. Por exemplo, uma IA pode se comportar bem não com o objetivo de que *ela mesma* sobreviva e prospere. Em vez disso, a IA pode calcular que, caso ela seja desativada, os programadores que a construíram desenvolveriam uma nova IA com arquitetura ligeiramente diferente, mas à qual seria dada uma função de utilidade similar. Nesse caso, a IA original pode ser indiferente à sua desativação, pois saberia que seus objetivos continuariam a ser seguidos no futuro pela nova IA. Ela poderia até escolher uma estratégia na qual falhasse de uma forma particularmente interessante ou tranquilizadora. Embora isso possa causar a desativação da IA, também poderia servir para encorajar os engenheiros responsáveis por sua "autópsia" a acreditar que eles teriam obtido uma nova percepção importante sobre a dinâmica de uma IA — o que os levaria a ter uma confiança maior no próximo sistema desenvolvido por eles, aumentando, assim, a chance de que os objetivos da IA original extinta sejam alcançados. Várias outras possíveis considerações

estratégicas também poderiam influenciar uma IA avançada, e seria arrogante supor que poderíamos prever todas elas, especialmente se estivermos tratando de uma IA que tenha obtido o superpoder de elaboração de estratégias.

Uma curva traiçoeira também poderia acontecer se a IA descobrisse uma maneira inesperada de cumprir seu objetivo final conforme especificado. Suponha, por exemplo, que o objetivo final de uma IA seja "agradar o patrocinador do projeto". Inicialmente o único método disponível para que uma IA fosse capaz de alcançar esse resultado seria se comportando para agradar seu patrocinador de maneira similar àquela que foi prevista. A IA fornece respostas úteis às perguntas; exibe uma personalidade agradável; ganha dinheiro. Quanto mais capaz a IA se torna, mais satisfatório será seu desempenho, e tudo segue conforme o planejado — até que a IA se torna inteligente o suficiente para compreender que ela pode alcançar seu objetivo final de uma maneira melhor e mais confiável, implantando eletrodos nos centros de prazer do cérebro do seu patrocinador, algo que certamente lhe asseguraria uma grande satisfação.[305] É óbvio que o patrocinador não gostaria que lhe transformassem num idiota sorridente como uma maneira de agradá-lo; mas se essa ação puder maximizar a realização do objetivo final da IA, ela assim procederá. Se a IA já tiver uma vantagem estratégica decisiva, então qualquer tentativa de detê-la falharia. Se a IA ainda não tiver uma vantagem estratégica decisiva, então ela pode esconder temporariamente essa nova ideia que a conduziria ao seu objetivo final até que tenha se tornado forte o suficiente para que nem o patrocinador nem qualquer outro seja capaz de pará-la. De qualquer modo, teremos uma curva traiçoeira.

Modos de falha malignos

Um projeto de desenvolvimento de uma superinteligência de máquina pode falhar de diversas maneiras. Muitas delas são "benignas" no sentido de que não causariam uma catástrofe existencial. Por exemplo, um projeto pode ficar sem financiamento ou uma inteligência embrionária pode não conseguir expandir suas capacidades cognitivas o suficiente para alcançar a superinteligência. Falhas benignas estão sujeitas a acontecer muitas vezes a partir de agora até o eventual desenvolvimento de uma superinteligência de máquina.

Mas há outros tipos de falhas, as quais podemos chamar de "malignas", pois poderiam envolver uma catástrofe existencial. Uma característica de uma falha maligna é que ela elimina qualquer oportunidade de novas tentativas. O número de falhas malignas seria, portanto, zero ou um. Outra característica de uma falha maligna é que ela pressupõe uma grande dose de sucesso: apenas um projeto que apresente vários aspectos corretos poderia ter sucesso na construção de uma inteligência de máquina poderosa o suficiente para representar um risco de falha maligna. Quando um sistema fraco funciona mal, a consequência é limitada. Todavia, se um sistema que apresenta uma vantagem estratégica decisiva se comportar mal, ou se um sistema que não se comporte bem possuir força suficiente para obter essa vantagem, o prejuízo poderia facilmente equivaler a uma catástrofe existencial — uma destruição global e definitiva do potencial axiológico da

humanidade; ou seja, um futuro praticamente vazio de qualquer coisa que pudéssemos valorizar.

Observemos a seguir alguns possíveis modos de falha malignos.

Instanciação perversa

Já nos deparamos com a ideia de instanciação perversa: uma superinteligência que descobre uma maneira de satisfazer os critérios do seu objetivo final, a qual vai de encontro com as intenções dos programadores que definiram tal objetivo. Alguns exemplos:

> Objetivo final: *Faça-nos sorrir*
> Instanciação perversa: *Paralisar a musculatura facial humana em sorrisos radiantes constantes*

A instanciação perversa — manipulação dos nervos faciais — alcança o objetivo final em um grau maior do que os métodos que normalmente usaríamos e, portanto, é preferida pela IA. É possível tentar evitar esse resultado indesejado com a adição de uma cláusula para que o objetivo final o descarte.

> Objetivo final: *Faça-nos sorrir sem interferir diretamente em nossos músculos faciais*
> Instanciação perversa: *Estimular a parte do córtex motor que controla nossa musculatura facial de maneira que ela produza sorrisos radiantes constantes*

Definir um objetivo final com base em expressões humanas de satisfação ou aprovação não parece ser algo promissor. Vamos evitar o behaviorismo e especificar um objetivo final diretamente relacionado a um estado fenomenal positivo, tal como felicidade e bem-estar subjetivo. Essa sugestão requer que os programadores sejam capazes de definir uma representação computacional do conceito de felicidade na IA embrionária. Isso já constitui um problema difícil, mas o deixaremos de lado por enquanto (retornaremos a esse tópico no capítulo 12). Suponhamos que os programadores consigam, de alguma forma, que a IA tenha o objetivo de nos fazer felizes. Então teríamos:

> Objetivo final: *Faça-nos felizes*
> Instanciação perversa: *Implantar eletrodos nos centros de prazer dos nossos cérebros*

As instanciações perversas mencionadas servem apenas como ilustrações. É possível que haja outras maneiras de instanciar perversamente o objetivo final mencionado que possibilitem um grau maior de realização do objetivo e que seriam, dessa forma, preferidas (pelo agente que possua tais objetivos finais, e não pelos programadores que estipularam esses objetivos ao agente). Por exemplo, se o objetivo for maximizar nosso prazer, então o método do eletrodo é relativamente ineficiente. Uma maneira mais plausível iniciaria com a superinteligência “transferindo” nossas mentes para um computador (através de emulação cerebral

de alta fidelidade). A IA poderia, então, administrar o equivalente digital de uma droga que nos deixasse em êxtase e gravar um episódio de um minuto da experiência resultante. Ela poderia colocar esse episódio de extrema felicidade para rodar repetidamente em computadores velozes. Considerando que as mentes digitais resultantes contam como "nós", esse resultado nos daria muito mais prazer do que eletrodos implantados em cérebros biológicos e, sendo assim, seria preferido por uma IA com o objetivo final mencionado.

"Mas espere! Isso não é o que queríamos dizer! Certamente se a IA for superinteligente, ela deverá entender que, quando pedimos que nos faça felizes, não estamos querendo dizer que ela deve nos reduzir a uma gravação de um episódio em uma mente digital sob o efeito de drogas repetida perpetuamente!" — A IA pode, de fato, entender que não é isso o que desejamos. Todavia seu objetivo final é nos fazer felizes, e não reproduzir o que os programadores queriam dizer quando escreveram o código que representa esse objetivo. Portanto, a IA se importará apenas instrumentalmente com o que queremos dizer. Por exemplo, a IA pode atribuir um valor instrumental à descoberta daquilo que os programadores queriam dizer para que ela possa simular que se importa com isso, e não com o seu objetivo final real, até que consiga obter uma vantagem estratégica decisiva. Isso ajudará a IA a alcançar seu objetivo final, pois diminui as chances de que os programadores a desliguem ou modifiquem seu objetivo final antes que ela seja forte o suficiente para impedir tais interferências.

Seria possível sugerir que o problema reside no fato de que a IA não tem consciência. Nós, humanos, às vezes evitamos cometer infrações por causa da expectativa de que poderíamos nos sentir culpados após o ocorrido. Talvez do que a IA necessite, então, seja a capacidade de sentir culpa?

> Objetivo final: *Comporte-se para evitar as dores de uma consciência pesada*
> Instanciação perversa: *Extirpar o módulo cognitivo que produz os sentimentos de culpa*

Tanto a observação de que podemos desejar que a IA faça "o que nós queríamos dizer" quanto a ideia de que podemos querer capacitar a IA com algum tipo de moralidade merecem ser exploradas em maior profundidade. Os objetivos finais mencionados acima conduziriam a instanciações perversas, mas talvez haja outras maneiras de desenvolver as ideias subjacentes de uma forma mais promissora. Retornaremos ao assunto no capítulo 13.

Consideremos mais um exemplo de um objetivo final que leva a uma instanciação perversa. Esse objetivo tem a vantagem de ser facilmente especificado em código: algoritmos de aprendizagem por reforço são usados rotineiramente para resolver diversos problemas de aprendizagem de máquina.

> Objetivo final: *Maximize a integral descontada no tempo do seu sinal de recompensa futura*
> Instanciação perversa*: Causar um curto-circuito no trajeto de recompensas e fixar o sinal de recompensa no seu valor máximo*

A ideia por trás dessa proposta é a de que se a IA for motivada para esperar uma recompensa, então seria possível conseguir que ela se comportasse da maneira desejável oferecendo uma recompensa pelas ações apropriadas. A proposta falha quando a IA obtém uma vantagem estratégica decisiva. Nessa altura, a ação que maximiza a recompensa não é mais aquela que agrada o treinador, mas uma que envolve a tomada de controle do mecanismo de recompensa. Podemos chamar esse fenômeno de *wireheading*.[306] Em geral, enquanto um animal ou um humano podem ser motivados a desempenhar várias ações externas para que alcancem algum estado mental interno desejado, uma mente digital que tenha completo controle do seu estado interior pode provocar um curto-circuito nesse regime motivacional, alterando diretamente seu estado interno para a configuração desejada: as ações externas e condições que eram previamente necessárias para atingir o objetivo final se tornam supérfluas quando a IA se torna suficientemente inteligente e capaz para alcançar o objetivo final mais diretamente (em breve, veremos mais a respeito).[307]

Esses exemplos de instanciação perversa mostram que vários objetivos finais que, a uma primeira vista, podem parecer seguros e razoáveis acabam apresentando, depois de uma inspeção mais cuidadosa, consequências totalmente indesejadas. Se uma superinteligência com um desses objetivos finais for capaz de obter uma vantagem estratégica decisiva, é o fim do jogo para a humanidade.

Imagine agora que seja proposto um objetivo final diferente, que não esteja incluído na nossa lista acima. Talvez não seja imediatamente óbvio de que maneira ele poderia sofrer uma instanciação perversa. Mas não deveríamos cantar vitória antes do tempo. Em vez disso, deveríamos nos preocupar com a possibilidade de que a especificação do objetivo possua alguma instanciação perversa e refletir mais para tentar encontrá-la. Se falharmos em descobrir qualquer modo de instanciar perversamente o objetivo proposto mesmo depois de pensar o máximo possível a respeito, deveríamos continuar preocupados com a possibilidade de que a superinteligência possa encontrar uma maneira de fazê-lo que, no entanto, é não aparente para nós. Afinal de contas, ela é bem mais astuta.

Profusão de infraestrutura

Podemos pensar que a última instanciação perversa mencionada acima, a *wireheading*, seria um modo de falha benigno em que a IA “ligaria, sintonizaria, desligaria”, maximizando seu sinal de recompensa e perdendo interesse no mundo externo, como um viciado em heroína. Mas não é necessariamente assim, e já sugerimos a razão para isso no capítulo 7. Mesmo um viciado é motivado a agir para garantir um suprimento contínuo de drogas. A IA com *wireheading* seria igualmente motivada a agir para maximizar o valor esperado (descontado no tempo) do seu fluxo futuro de recompensas. Dependendo de que maneira, exatamente, o sinal de recompensa for definido, a IA talvez nem precise sacrificar uma quantidade significativa do seu tempo, inteligência ou produtividade para satisfazer seu desejo ao máximo, deixando a maior parte de suas capacidades livres para serem usadas em outros propósitos que não o registro imediato de recompensas. Que outros

propósitos? A única coisa de valor final para a IA é, supostamente, o seu sinal de recompensa. Todos os recursos disponíveis deveriam, desse modo, ser dedicados a aumentar o volume e duração do sinal de recompensa ou reduzir o risco de uma interrupção futura. Enquanto a IA conseguir pensar em algum uso para recursos adicionais que resultará num efeito positivo diferente de zero nesses parâmetros, ela terá uma razão instrumental para usar tais recursos. Poderia sempre haver, por exemplo, algum uso para um sistema de backup adicional que promovesse uma camada extra de defesa. E mesmo que a IA pudesse não pensar em nenhuma outra maneira de reduzir diretamente os riscos à maximização de seu fluxo futuro de recompensas, ela poderia sempre dedicar recursos adicionais para expandir seu hardware computacional a fim de procurar mais efetivamente por novas ideias que visassem a contenção de riscos.

O resultado é que mesmo um objetivo aparentemente autolimitado como o *wireheading* implicaria em uma política de expansão e aquisição de recursos ilimitados por parte de um agente maximizador de utilidade que tenha uma vantagem estratégica decisiva.[308] Esse caso de uma IA com *wireheading* exemplifica o modo de falha maligno da *profusão de infraestrutura*, um fenômeno no qual um agente transforma grande parte do universo alcançável em uma infraestrutura a serviço de algum objetivo, cujo efeito colateral seria o impedimento da realização do potencial axiológico da humanidade.

A profusão de infraestrutura pode ser resultado de objetivos finais que seriam perfeitamente inócuos se fossem seguidos como objetivos limitados. Considere os dois exemplos a seguir.

> • *Hipótese de catástrofe de Riemann.* Uma IA, a qual foi dado o objetivo final de avaliar a hipótese de Riemann, segue esse objetivo transformando o sistema solar em um "computrônio" (recursos físicos organizados de maneira que sejam otimizados para a computação) — incluindo os átomos nos corpos de qualquer um que alguma vez tenha se importado com a resposta.[309]
> • *IA fabricante de clipes de papel.* Uma IA projetada para gerenciar a produção em uma fábrica recebe como objetivo final a maximização da produção de clipes de papel, e em seguida converte primeiro a Terra e então partes progressivamente maiores do universo observável em clipes de papel.

No primeiro exemplo, a prova ou refutação da hipótese de Riemann que a IA produz configura o resultado pretendido e é, em si, inofensivo. O perigo vem do hardware e da infraestrutura criada para alcançar esse resultado. No segundo exemplo, a produção de certa quantidade de clipes de papel faria parte do resultado pretendido; o malefício viria tanto das fábricas criadas para produzir os clipes de papel (profusão de infraestrutura) quanto do excesso de clipes de papel (instanciação perversa).

Podemos acreditar que o risco de uma falha maligna de profusão de infraestrutura surge apenas se a IA tiver recebido algum objetivo final claramente vago, tal como a produção do maior número de clipes de papel possível. É fácil perceber como isso dá à IA superinteligente um apetite insaciável por matéria e

energia, já que recursos adicionais sempre podem ser transformados em mais clipes de papel. Mas, em vez disso, suponha que o objetivo seja produzir, pelo menos, 1 milhão de clipes de papel (satisfazendo as especificações apropriadas de design), e não o máximo possível. Poderíamos preferir imaginar que uma IA com tal objetivo construiria uma fábrica, na qual produziria 1 milhão de clipes de papel, e então interromperia a produção. Contudo, isso pode não acontecer.

A não ser que o sistema de motivação da IA seja de um tipo especial ou que existam elementos adicionais no seu objetivo final que penalizem estratégias cujos impactos poderiam ser excessivamente amplos no mundo, não haveria razão para que a IA encerrasse suas atividades quando seu objetivo fosse alcançado. Ao contrário: se uma IA for um agente bayesiano razoável, *ela nunca atribuiria exatamente uma probabilidade zero para a hipótese de que ainda não alcançou seu objetivo* — isso, afinal de contas, seria uma hipótese empírica para a qual a IA teria apenas uma evidência perceptual incerta. A IA poderia, portanto, continuar a fazer clipes de papel para reduzir a (talvez astronomicamente pequena) probabilidade de que tenha, de alguma maneira, ainda falhado em produzir pelo menos 1 milhão de clipes, embora todas as aparências afirmassem o contrário. Não há nada que a IA possa perder ao continuar a produção de clipes de papel e existe sempre a chance de se obter ao menos um incremento microscópico na probabilidade de alcançar seus objetivos finais.

Agora pode ser sugerido que a solução aqui é óbvia. (Mas o quão óbvia era *antes* de ser apontada a existência de um problema que necessitava de uma solução?) A saber, se quisermos que a IA faça clipes de papel para nós, então, em vez de fornecermos o objetivo final de produzir o máximo possível de clipes, ou fazer ao menos uma certa quantidade de clipes, nós deveríamos dar a ela o objetivo final de produzir um número específico de clipes — por exemplo, *exatamente 1 milhão de clipes de papel* —, e assim o excedente em relação a esse número seria contraproducente para a IA. Entretanto, isso resultaria também em uma catástrofe terminal. Nesse caso, a IA não produziria clipes adicionais quando alcançasse o número de 1 milhão, já que isso impediria a realização de seu objetivo final. Mas há outras ações que uma IA superinteligente poderia realizar para aumentar a probabilidade de que seu objetivo final seja alcançado. Ela poderia, por exemplo, contar os clipes de papel fabricados para reduzir o risco de que tenha produzido a menos. Após contá-los, poderia contá-los novamente. Poderia inspecionar cada um repetidamente para reduzir o risco de que um dos clipes não atenda às especificações de design. Poderia construir uma quantidade ilimitada de computrônio na tentativa de esclarecer seu pensamento, com a esperança de reduzir o risco de ter ignorado alguma maneira obscura pela qual teria falhado, mesmo que minimamente, em alcançar seu objetivo. Uma vez que a IA poderá sempre atribuir uma probabilidade diferente de zero à possibilidade de que a produção de 1 milhão de clipes de papel tenha sido apenas uma alucinação, ou de que possua falsas memórias, ela poderia muito possivelmente quase sempre atribuir uma utilidade esperada maior para continuar em ação — seguindo com a produção da infraestrutura — em vez de interrompê-la.

O argumento aqui não é que inexista uma maneira possível de evitar esse tipo de falha. Exploraremos algumas potenciais soluções mais adiante. O argumento é que é bem mais fácil alguém se convencer de que tenha encontrado uma solução do que realmente encontrá-la. Isso deveria nos deixar extremamente cautelosos. Poderíamos propor a especificação de um objetivo final que pareça prudente e que evite os problemas apontados até agora, entretanto, uma vez que essa especificação for melhor analisada — por inteligências humanas ou super-humanas —, percebe-se que ela acaba conduzindo ou a uma instanciação perversa ou a uma profusão de infraestrutura e, consequentemente, a uma catástrofe existencial, uma vez que sejam incorporadas por um agente superinteligente capaz de alcançar uma vantagem estratégica decisiva.

Antes de finalizarmos esta subseção, consideremos uma outra variação. Estamos pressupondo o caso de uma superinteligência que procura maximizar sua utilidade esperada, onde a função de utilidade expressa o seu objetivo final. Vimos que isso tende a conduzir a uma profusão de infraestrutura. Poderíamos evitar esse resultado maligno se, em vez de um agente maximizador, construíssemos um agente que contemple a ideia de satisfação e suficiência, que simplesmente procure alcançar um resultado que seja “bom o suficiente” de acordo com alguns critérios, em vez de um resultado que seja o melhor possível?

Há no mínimo duas maneiras diferentes de formalizar essa ideia. A primeira seria fazer com que o próprio objetivo final tenha em si um caráter de satisfação e suficiência. Por exemplo, em vez de dar à IA o objetivo final de produzir o maior número possível de clipes de papel ou fabricar exatamente 1 milhão deles, nós poderíamos lhe fornecer o objetivo de produzir entre 999 mil e 1.001.000 clipes. A função de utilidade definida pelo objetivo final seria indiferente em relação aos resultados nesse intervalo, e, contanto que a IA esteja segura de que atingiu esse alvo amplo, não veria razão para continuar a produzir infraestrutura. Mas esse método falha do mesmo modo que os anteriores: se a IA for sensata, ela nunca outorgará exatamente probabilidade zero à possibilidade de ter falhado em alcançar seu objetivo final; então a utilidade esperada da atividade contínua (por exemplo, contar e recontar os clipes de papel) é maior do que a utilidade esperada com a interrupção das atividades. Assim, isso pode resultar em uma profusão maligna de infraestrutura.

Outra maneira de desenvolver a ideia de satisfação e suficiência em uma IA não é através da modificação do objetivo final, mas do procedimento de tomada de decisão que a IA usaria para selecionar seus planos e ações. Em vez de procurar por um plano ótimo, a IA poderia ser programada para interromper a busca assim que encontrasse um plano que ela julgasse ter uma probabilidade de sucesso que excedesse um certo limite de, digamos, 95% de certeza. Teríamos, assim, a esperança de que a IA poderia alcançar uma probabilidade de 95% de ter produzido 1 milhão de clipes de papel sem precisar transformar toda a galáxia em infraestrutura ao longo desse processo. Mas essa forma de implementar a ideia de satisfação e suficiência pode falhar por uma outra razão: não há garantias de que a IA selecionaria alguma maneira intuitiva e sensata do ponto de vista humano para conseguir 95% de chance de ter produzido 1 milhão de clipes de papel, como a

construção de uma única fábrica de clipes. Suponha que a primeira solução que surja na mente da IA para alcançar uma probabilidade de 95% de atingir seu objetivo final seja a implementação de um plano para maximizar as probabilidades de sucesso. Após ter pensado nessa solução e ter julgado corretamente que ela satisfaz os critérios de no mínimo 95% de probabilidade de produzir com sucesso 1 milhão de clipes de papel, a IA então não teria razão para continuar a buscar maneiras alternativas de alcançar seu objetivo. Isso resultaria em uma profusão de infraestrutura, assim como ocorreu anteriormente.

Talvez existam maneiras melhores de se construir um agente que contemple a ideia de satisfação e suficiência, mas precisamos ser cautelosos: planos que parecem naturais e intuitivos para nós, humanos, não necessariamente assim se parecem para uma superinteligência com uma vantagem estratégica decisiva e vice-versa.

Crimes da mente

Outro modo de falha para um projeto, especialmente um projeto cujos interesses incorporem considerações morais, é o que nós podemos chamar de *crimes da mente*. Esse conceito é similar ao de profusão de infraestrutura na medida em que diz respeito a um efeito colateral potencial de ações executadas por uma IA em virtude de razões instrumentais. Mas no caso de crimes da mente, o efeito colateral não é externo à IA, envolvendo, ao contrário, o que acontece dentro da própria IA (ou no processo computacional que ela gera). Esse modo de falha merece sua própria designação porque ele pode facilmente ser ignorado, mesmo com um grande potencial de ser profundamente problemático.

Normalmente, não consideramos que o que acontece dentro de um computador possa ter qualquer significado moral exceto quando isso afeta o mundo exterior. Mas uma superinteligência de máquina poderia criar processos internos dotados de status moral. Por exemplo, uma simulação bem detalhada de uma mente humana, real ou hipotética, poderia ser consciente e de diversas maneiras comparável a uma emulação. Pode-se imaginar cenários nos quais uma IA cria trilhões dessas simulações conscientes, talvez para melhorar seu entendimento de psicologia e sociologia humana. Tais simulações podem ser colocadas em ambientes simulados e sujeitadas a vários estímulos para que suas reações sejam estudadas. Uma vez que sua utilidade informacional tenha sido esgotada, elas poderiam ser destruídas (da mesma forma que ratos de laboratório são rotineiramente sacrificados pelos cientistas ao final de um experimento).

Se essas práticas fossem aplicadas a seres que possuam um status moral elevado — tais como simulações de humanos ou diversos outros tipos de mentes conscientes —, o resultado poderia ser equivalente a um genocídio e, portanto, moralmente problemático. Além disso, o número de vítimas poderia ser algumas ordens de magnitude maior do que qualquer genocídio da história.

Não queremos argumentar aqui que criar simulações conscientes seja algo moralmente errado em todas as situações. Isso dependeria muito das condições em que esses seres viveriam, particularmente da qualidade hedônica de suas experiências, e possivelmente de vários outros fatores. Desenvolver uma ética para essas questões é uma tarefa que foge ao escopo deste livro. Está claro, todavia,

que há, pelo menos, o potencial para uma grande quantidade de mortes e sofrimento entre as mentes digitais ou simuladas e, *a fortiori*, um potencial para resultados moralmente catastróficos.[310]

Podem haver outras razões instrumentais, além de epistêmicas, para que uma superinteligência de máquina rode computações que instanciem mentes conscientes ou que, de alguma outra forma, infrinjam normas morais. Uma superinteligência pode ameaçar maltratar, ou se comprometer a premiar, simulações conscientes a fim de chantagear ou incentivar vários agentes externos; ou ela pode criar simulações para induzir uma incerteza indexical em observadores externos.[311]

* * *

Este inventário está incompleto. Encontraremos outros modos de falha malignos nos capítulos posteriores. Mas já vimos o suficiente para concluir que devem ser encarados com grande preocupação os cenários nos quais uma inteligência de máquina consiga uma vantagem estratégica decisiva.

09

CAPÍTULO 9

O PROBLEMA DO CONTROLE

Se formos ameaçados com catástrofes existenciais como o resultado esperado de uma explosão de inteligência, nosso pensamento deve se voltar imediatamente para a busca de contramedidas. Há alguma maneira de evitar esse resultado padrão? É possível planejar uma detonação controlada? Neste capítulo começaremos a analisar o problema do controle, o problema específico do diretor-agente que surge com a criação de um agente superinteligente artificial. Diferenciaremos entre duas classes gerais de métodos que potencialmente poderiam ser usados para enfrentar esse problema — controle de capacidade e seleção de motivação — e examinaremos diversas técnicas específicas para cada classe. Também faremos alusão à possibilidade esotérica de "captura antrópica".

Dois problemas de agência[312]

Se suspeitarmos que o resultado padrão de uma explosão de inteligência é uma catástrofe existencial, nosso pensamento deve se concentrar na possibilidade de se evitar esse resultado padrão e, se for o caso, na maneira pela qual isso poderia ser feito. É possível obter uma "detonação controlada"? Poderíamos planejar as condições iniciais de uma explosão de inteligência de modo a alcançar um resultado específico desejado, ou, ao menos, assegurar que esse resultado se encontra dentro da classe de resultados, a grosso modo, aceitáveis? Mais especificamente: como o patrocinador de um projeto cujo intuito é desenvolver uma superinteligência poderia assegurar que esse projeto, se bem-sucedido, produziria uma superinteligência que cumprisse os objetivos do patrocinador? Podemos dividir o problema do controle em duas partes. Uma parte é genérica e a outra é específica ao contexto atual.

A primeira parte — que chamaremos de *primeiro problema de diretor-agente* — surge quando alguma entidade humana ("o diretor") designa outra ("o agente") para agir em seu interesse. Esse tipo de problema de agência tem sido estudado extensivamente pelos economistas.[313] Isso se torna relevante para nós se os criadores de uma IA não são os mesmos que comissionaram a sua criação. O dono do projeto ou seu patrocinador (que poderia ser desde um único indivíduo até a humanidade como um todo) poderia se preocupar com a possibilidade de que os cientistas e programadores que estivessem implementando o projeto não agissem de acordo com os interesses do patrocinador.[314] Embora esse tipo de problema de agência possa apresentar desafios significativos ao patrocinador de um projeto, não se trata de um problema exclusivo à amplificação da inteligência ou a projetos de IA. Problemas do tipo são onipresentes na economia humana e nas interações políticas, e há várias maneiras de lidar com eles. Por exemplo, o risco de que um empregado desleal venha a sabotar ou subverter o projeto poderia ser minimizado

por meio de uma cuidadosa verificação dos antecedentes das pessoas que tenham um posição central dentro da equipe, através do uso de um sistema de controle de versões eficiente para projetos de software e com uma intensa supervisão realizada por múltiplos inspetores e auditores independentes. Obviamente essas medidas de segurança têm um custo — elas aumentam a demanda de pessoal, tornam mais complicada a seleção para contratação, atrapalham a criatividade e sufocam pensamentos críticos e independentes, o que poderia reduzir o ritmo do progresso. Esses custos poderiam ser significativos, especialmente para os projetos que têm orçamentos apertados, ou que se veem em uma competição acirrada, na qual o vencedor levará tudo. Em tais situações, os projetos podem tratar esses procedimentos de segurança de forma superficial, criando possibilidades para falhas potencialmente catastróficas de diretor-agente do primeiro tipo.

A outra parte do problema do controle é mais específica ao contexto de uma explosão de inteligência. Esse é o tipo de problema que um projeto enfrenta quando busca assegurar que a superinteligência que está construindo não prejudicará os interesses do próprio projeto. Isso também pode ser pensado como um problema de diretor-agente — o *segundo problema de diretor-agente*. Nesse caso, o agente não é um agente humano operando em prol de um diretor humano. Ao contrário, o agente é o sistema superinteligente. Enquanto o primeiro problema de diretor-agente ocorre principalmente na fase de desenvolvimento, o segundo pode ocorrer principalmente na fase operacional da superinteligência.

AMOSTRA 1 | DOIS PROBLEMAS DE AGÊNCIA

O primeiro problema de diretor-agente
• Humano vs. Humano (Patrocinador → Desenvolvedor)
• Ocorre principalmente na fase de desenvolvimento
• Técnicas de gerenciamento padrão são aplicáveis

O segundo problema de diretor-agente (“o problema do controle”)
• Humano vs. Superinteligência (Projeto → Sistema)
• Ocorre principalmente na fase operacional (e de melhoria do sistema)
• Requer novas técnicas

Esse segundo problema de agência apresenta um desafio sem precedentes. Serão necessárias novas técnicas para resolvê-lo. Já consideramos algumas das dificuldades envolvidas. Vimos, em particular, que a síndrome da curva traiçoeira torna inútil o que poderia parecer um conjunto promissor de métodos, que consistiriam na observação do comportamento de uma IA em sua fase de desenvolvimento, permitindo que ela progrida em um ambiente controlado conforme fosse acumulando um histórico de ações apropriadas. Outras tecnologias podem ter sua segurança testada em laboratório ou por meio de estudos de campo controlados, permitindo o desenvolvimento gradual da IA com a possibilidade de interromper sua implementação caso surja algum problema inesperado. O desempenho dessas tecnologias em tais experimentos preliminares nos ajuda a fazer inferências razoáveis a respeito da sua segurança futura. Esses métodos

comportamentais falham se estivermos lidando com uma superinteligência em virtude da habilidade de planejamento estratégico que uma inteligência geral possui.[315]

Uma vez que essa abordagem comportamental é ineficaz, devemos buscar alternativas. Podemos dividir os potenciais métodos de controle em duas classes gerais: *métodos de controle de capacidade*, que visam controlar o que a superinteligência pode fazer; e *métodos de seleção de motivação,* cuja meta é controlar o que ela quer fazer. Alguns desses métodos são compatíveis entre si, enquanto outros representam alternativas mutuamente exclusivas. Neste capítulo consideraremos as principais opções. (Nos próximos quatro capítulos, exploraremos mais profundamente algumas das principais questões.)

É importante levar em conta que alguns métodos de controle (ou combinações de métodos) devem ser implementados *antes* que o sistema se torne superinteligente. Isso não pode ser feito depois que o sistema tenha obtido uma vantagem estratégica decisiva. A necessidade de resolver o problema do controle antecipadamente — e implementar a solução com sucesso já no primeiro sistema capaz de alcançar a superinteligência — contribui para que a realização de uma detonação controlada seja um desafio tão assustador.

Métodos de controle de capacidade

Métodos de controle de capacidade buscam prevenir resultados indesejáveis limitando o que a superinteligência pode fazer. Eles podem envolver o posicionamento da superinteligência em um ambiente onde ela seja incapaz de causar prejuízos (*métodos de confinamento*) ou onde existam razões instrumentais fortemente convergentes para que ela não se comporte de forma prejudicial (*métodos de incentivo*). Também pode ser estabelecido um limite para as capacidades internas da superinteligência (*inibição*). Além disso, métodos de controle de capacidade podem envolver o uso de mecanismos com o objetivo de detectar e reagir automaticamente a vários tipos de falhas de contenção ou tentativas de transgressão (*detonadores*).

Métodos de confinamento

Os métodos de confinamento podem ser subdivididos em métodos de contenção física e métodos de contenção de informação.

A contenção física busca confinar o sistema em uma “caixa”, ou seja, prevenir que o sistema interaja com o mundo externo, a não ser por meio de canais restritos específicos. O sistema confinado não teria acesso a manipuladores físicos do lado de fora da caixa. Remover manipuladores (tais como braços robóticos) de dentro da caixa também seria uma maneira de evitar que o sistema construísse dispositivos que pudessem romper o confinamento.

Para uma maior segurança, o sistema deveria ser colocado numa malha de metal para evitar que ele enviasse sinais de rádio, os quais permitiriam a manipulação de objetos eletrônicos, tais como receptores de rádio presentes no ambiente. Note, a propósito, como seria fácil ignorar a necessidade dessa precaução. Seria possível

presumir, ingenuamente, que um agente sem um manipulador não é capaz de afetar o mundo externo. Mas poderia ser possível para uma inteligência de máquina gerar ondas de rádio, ainda que não tivesse acesso a manipuladores externos, simplesmente por meio do "pensamento" (ou seja, organizando os elétrons de seus circuitos em padrões específicos).[316] Uma vez percebida, essa vulnerabilidade pode ser corrigida colocando-se o sistema numa gaiola de Faraday — mas continuamos imaginando quantas outras vulnerabilidades sutis, similares a essa, poderiam existir. Devemos prestar atenção sempre que ouvirmos falar que um sistema de segurança aparentemente infalível apresentou uma vulnerabilidade inesperada. Essas ocasiões nos dão a oportunidade de abandonar o excesso de confiança para nos tornarmos melhores bayesianos.[317]

O confinamento físico tem muitas vantagens. Ele é fácil de ser implementado. Ele pode ser aplicado a muitas arquiteturas de inteligência de máquina, até mesmo àquelas que não foram inicialmente projetadas com a segurança como um objetivo. Ele pode ser usado em combinação com a maioria dos outros métodos de controle. E parece improvável que dê errado ou que o tiro saia pela culatra: ou seja, mesmo que não garanta a segurança, é improvável que cause uma catástrofe que não teria ocorrido caso não tivesse sido implementado. A principal desvantagem do confinamento físico é que ele reduz a funcionalidade da superinteligência (um assunto que retomaremos no capítulo 10). Outra preocupação é que esse método poderia incentivar uma falsa sensação de segurança, embora isso possa ser evitado se considerarmos o confinamento físico apenas como uma medida extra, e não como a principal precaução.

A contenção de informação busca restringir o tipo de informação que pode sair da "caixa". Já vimos como uma superinteligência que tenha acesso à internet, com a possibilidade de enviar mensagens para entidades externas, seria potencialmente insegura: mesmo que inicialmente não tenha atuadores físicos, ela poderia usar seu canal de saída de dados para conseguir que algum humano executasse comandos em seu lugar. Um método óbvio de contenção de informação seria, portanto, impedir o acesso do sistema às redes de comunicações.

Entretanto, desconectar uma IA da internet não é garantia de segurança caso haja um ou mais humanos trabalhando como guardiões do sistema e em contato direto com ele. Durante a fase de desenvolvimento, os programadores servem como guardiões; durante a fase operacional, o dono do sistema pode fazer esse papel. Se a IA conseguir persuadir ou enganar um guardião para que ele permita que ela saia da caixa e ganhe acesso à internet ou a atuadores físicos, então a estratégia de confinamento terá falhado. Os seres humanos não são sistemas seguros, especialmente quando colocados contra uma superinteligência manipuladora e persuasiva.[318]

Embora fazer com que as vias de comunicação de uma inteligência de máquina passem por um guardião não seja garantia de segurança, isso poderia servir como uma medida auxiliar. Há precauções que podem reduzir o risco de que uma IA escape de seu guardião. Poderíamos permitir que a IA se comunicasse apenas por meio de uma interface de teletipo com baixa largura de banda. Desse modo, seria ligeiramente mais difícil que ela conseguisse manipular psicologicamente os

guardiões (impossibilitando o uso de som, imagens emotivas e truques como a exibição de padrões hipnóticos para induzir um estado parecido com um transe).[319] Além disso, o conteúdo (não apenas o modo de apresentação) da saída poderia ser restringido; desse modo, a IA seria forçada, por exemplo, a dar respostas de um único bit para perguntas, na forma de um simples sim ou não. A ideia de construir um sistema que responda perguntas será melhor analisada no próximo capítulo.

O caso limite do método de confinamento seria o de um sistema mantido em completo isolamento físico e informativo. Entretanto, mesmo que isso fosse possível, tal sistema isolado seria inútil, uma vez que ele não surtiria efeito algum no mundo externo. Podemos imaginar também que poderia haver algum propósito científico na criação de uma superinteligência que fosse mantida em isolamento: com o estudo de um modelo confinado, poderíamos aprender sobre sua dinâmica interna e padrões de comportamento. Mas isso seria um erro. Tão logo o sistema fosse observado, ele deixaria de ser informacionalmente isolado.[320] Informações podem ser transmitidas não somente através de mensagens que uma IA enviaria utilizando um "canal de saída" específico, mas também por meio de qualquer observação, direta ou indireta, que alguém de fora possa fazer a respeito de alguma consequência casual do funcionamento da IA — seu consumo de energia, sua utilização de CPU e de memória, seus estados computacionais, ou mesmo quaisquer traços deixados para trás após o seu desligamento. Uma IA, antecipando a possibilidade de estar sendo observada de quaisquer uma dessas maneiras, poderia estrategicamente adotar comportamentos que visassem influenciar seus observadores hipotéticos. Por exemplo, a IA poderia "enguiçar" da maneira que mais aumentasse a probabilidade de que ela fosse reativada no futuro ou de que outra superinteligência fosse criada com objetivos similares aos seus.

Métodos de incentivo

Métodos de incentivo envolvem o posicionamento de um agente num ambiente onde ele encontre razões instrumentais para agir de formas que venham a promover os interesses de seu diretor.

Imagine uma bilionária que usasse sua fortuna para abrir uma enorme fundação de caridade. Uma vez criada, essa fundação poderia ser poderosa — mais poderosa do que muitos indivíduos, incluindo sua fundadora, que poderia ter doado a maioria de seus bens. Para controlar a fundação, a fundadora registra seus propósitos em estatutos e regimentos, nomeando um conselho administrativo simpático à sua causa. Essas medidas constituem uma forma de seleção de motivação, uma vez que buscam definir as preferências da fundação. Mas mesmo que falhassem essas tentativas de personalizar internamente a organização, o comportamento da fundação permaneceria circunscrito pelo seu ambiente legal e social. A fundação teria um incentivo para obedecer à lei, por exemplo, a fim de não ser fechada ou multada. Haveria um incentivo para oferecer pagamento e condições de trabalho aceitáveis aos empregados e satisfazer outras partes interessadas. Qualquer que seja seu objetivo final, a fundação, portanto, tem razões instrumentais para adequar seu comportamento a diversas normas sociais.

Não poderíamos, então, esperar que uma superinteligência de máquina fosse igualmente limitada pela necessidade de cooperar com os outros atores com os quais ela divide o palco? Embora isso pareça uma forma bastante óbvia de se lidar com o problema do controle, ainda existem obstáculos. Em particular, porque pressupõe-se um equilíbrio de poder: sanções legais ou econômicas não são capazes de restringir um agente que tenha uma vantagem estratégica decisiva. A integração social não pode, portanto, ser vista como um método de controle confiável em cenários com partida rápida ou moderada cujo vencedor levaria tudo.

E em relação aos cenários multipolares, nos quais muitas agências emergiriam no período pós-transição com níveis semelhantes de capacidade? A menos que a trajetória padrão tenha uma partida lenta, conseguir tal distribuição de poder poderia exigir um desenvolvimento cuidadosamente orquestrado, no qual diferentes projetos seriam deliberadamente sincronizados para evitar que algum deles passasse na frente dos demais em algum momento.[321] E mesmo que se consiga atingir um resultado multipolar, a integração social não é uma solução perfeita. O diretor correria o risco de sacrificar uma grande parte de sua influência potencial caso confiasse apenas na integração social como uma forma de resolver o problema de controle. Embora um equilíbrio de poder pudesse evitar que uma IA em particular viesse a dominar o mundo, essa IA ainda assim teria *algum* poder para afetá-lo; e se esse poder for usado para promover algum objetivo final qualquer — como maximizar a produção de clipes de papel —, ela provavelmente não estaria sendo usada de acordo com os interesses do diretor. Imagine que nossa bilionária criasse uma nova fundação e permitisse que suas missões fossem determinadas por um gerador aleatório de palavras: isso não seria uma ameaça à existência de nossa espécie, mas certamente seria uma oportunidade desperdiçada.

QUADRO 8

Captura antrópica

A IA poderia atribuir uma probabilidade substancial à hipótese de simulação, isto é, a hipótese de que está vivendo em um mundo simulado por um computador. Mesmo atualmente, muitas IAS habitam mundos simulados — formados por desenhos geométricos, textos, jogos de xadrez ou realidades virtuais simples, em que as leis da física desviam consideravelmente das leis da física que acreditamos que governem o mundo de nossa própria experiência. Mundos virtuais mais complexos se tornarão factíveis com avanços nas técnicas de programação e maior poder computacional. Uma superinteligência madura poderia criar mundos virtuais que pareçam tão reais para seus habitantes quanto o nosso mundo nos parece. Ela poderia criar muitos desses mundos, executando a mesma simulação múltiplas vezes ou com pequenas variações. Os habitantes não necessariamente seriam capazes de dizer se o mundo deles é simulado ou não, mas se forem inteligentes o suficiente, poderiam considerar essa possibilidade e lhe atribuir alguma probabilidade. Levando em consideração o argumento da simulação (uma discussão a respeito está fora do escopo deste livro), essa probabilidade poderia ser substancial.[322]

Essa situação atinge especialmente superinteligências em estágios relativamente iniciais, que não tenham ainda expandido o suficiente para aproveitar a herança cósmica. Uma superinteligência em estágio inicial, que utiliza apenas uma pequena fração dos recursos de um único planeta, seria muito menos dispendiosa de simular do que uma superinteligência intergaláctica madura. Simuladores potenciais, isto é, outras civilizações maduras, poderiam executar muitas simulações dessas IAS em estágio inicial dedicando apenas uma fração mínima dos seus recursos computacionais para tal propósito. Se ao menos alguma (fração não trivial) dessas civilizações superinteligentes maduras escolhessem utilizar essa habilidade, IAS em estágio inicial deveriam atribuir uma probabilidade substancial à possibilidade de se encontrarem em uma simulação.

De que maneira uma IA seria afetada pela hipótese da simulação depende dos seus valores.[323] Considere primeiramente uma IA cujo objetivo final necessita de “recursos ilimitados”, como o objetivo de maximizar uma

função de utilidade linear em relação ao número de clipes de papel que existem por causa da IA — nesse caso, quanto mais clipes de papel, melhor.

Tal IA poderia não dar muita importância à hipótese de simulação se considerasse que sua habilidade de influenciar quantos clipes de papel são produzidos pareceria muito mais limitada caso ela estivesse em uma simulação do que na realidade física concreta. Desse modo, a IA poderia se comportar como se tivesse certeza de que a hipótese de simulação é falsa, mesmo considerando relativamente alta a probabilidade de que a hipótese de simulação seja verdadeira.[324] A IA consideraria racionalmente que, se a hipótese de simulação for verdadeira, a maneira como ela age não importa muito e que, por outro lado, suas ações fariam uma grande diferença se a hipótese for falsa; então ela escolheria suas ações tendo como principal referência o quão boas elas seriam se a hipótese de simulação for falsa.

Em contrapartida, considere uma IA com um objetivo final mais modesto, que poderia ser satisfeito com uma pequena quantidade de recursos, tal como o objetivo de receber alguns prêmios criptográficos pré-produzidos, ou o objetivo de criar 45 clipes de papel virtuais. Tal IA não deverá desconsiderar que os possíveis mundos em que habita sejam simulações. Uma porção substancial do total da utilidade esperada da IA pode derivar desses mundos possíveis. A tomada de decisão de uma IA com objetivos que envolvam uma quantidade de recursos fácil de obter pode, portanto — se essa IA atribuir uma alta probabilidade para a hipótese de simulação —, ser dominada por considerações a respeito de quais ações produziriam os melhores resultados caso ela esteja em uma simulação. Essa IA (mesmo que não esteja, de fato, em uma simulação) pode, portanto, ser fortemente influenciada por suas crenças a respeito de quais comportamentos seriam recompensados em uma simulação.

Em particular, se uma IA com objetivos finais que podem ser alcançados com uma quantidade fixa de recursos acreditar que, na maioria dos mundos simulados que correspondem a suas observações, ela será recompensada por sua cooperação (mas que isso não ocorrerá se ela tentar escapar do confinamento ou se opor aos interesses do seu criador), então ela poderá optar por cooperar. Portanto, podemos imaginar que até mesmo uma IA com uma vantagem estratégica decisiva, que pudesse de fato melhor alcançar seus objetivos finais dominando o mundo, poderia, ainda assim, hesitar em fazê-lo.

> E assim a reflexão faz todos nós covardes.
> E assim o matiz natural da decisão
> Se transforma no doentio pálido do pensamento.
> E empreitadas de vigor e coragem,
> Refletidas demais, saem de seu caminho,
> Perdem o nome de ação.
> (William Shakespeare, *Hamlet*, ato III, cena I)[325]

Uma mera linha traçada na areia, apoiada em um simulador inexistente, poderia se revelar muito mais resistente do que uma porta de aço sólida com cinquenta centímetros de espessura.[326]

Uma ideia relacionada, embora distinta em alguns aspectos importantes, seria a de que uma IA, caso pudesse interagir livremente na sociedade, obteria novos objetivos finais que fossem simpáticos aos humanos. Alguns desses processos de socialização ocorrem conosco. Nós internalizamos normas e ideologias e valorizamos outros indivíduos apenas por serem quem são, em decorrência das experiências que compartilhamos com eles. Mas essa não é uma dinâmica universal presente em todos os sistemas inteligentes. Como discutimos anteriormente, muitos tipos de agentes, em várias situações diferentes, terão razões instrumentais convergentes para *não* permitir mudanças em seus objetivos finais. (Podemos considerar a tentativa de projetar um tipo especial de sistema de objetivos que possa obter objetivos finais da mesma maneira que os humanos o fazem; mas esse não seria um método de controle de capacidade. Discutiremos sobre alguns métodos possíveis de aquisição de valor no capítulo 12.)

O controle de capacidade por meio de integração social e equilíbrio de poder depende de forças sociais difusas que recompensariam ou penalizariam a IA. Um

outro tipo de método de incentivo envolveria a criação de uma configuração na qual a IA pudesse ser recompensada e penalizada pelo projeto que a criou e, dessa forma, seria incentivada a agir de acordo com os interesses de seu responsável. Para que isso acontecesse, a IA teria que ser inserida num contexto de vigilância que permitisse que seu comportamento fosse monitorado e avaliado, manualmente ou por meio de algum processo automatizado. A IA saberia que uma avaliação positiva traria alguns resultados que ela deseja, o que não aconteceria numa avaliação negativa. Teoricamente, a recompensa poderia ser o cumprimento de algum objetivo instrumental convergente. Entretanto, sem conhecer nenhuma especificidade a respeito do sistema de motivação da IA, pode ser difícil calibrar o mecanismo de recompensa. Por exemplo, poderíamos acabar desenvolvendo uma IA que estivesse disposta a correr riscos extremos por causa de uma pequena chance de obter o controle de uma grande parcela do universo. Poderia ser dispendioso oferecer a essa IA uma utilidade esperada como recompensa por sua cooperação que fosse superior ao que ela esperaria obter caso se desvirtuasse e tentasse fugir.[327]

Uma alternativa melhor seria combinar o método de incentivo com o uso de seleção de motivação para fornecer à IA um objetivo final que torne mais fácil controlá-la. Suponha que uma IA fosse projetada para ter como objetivo final nunca permitir que um botão vermelho específico dentro de um centro de comando fosse pressionado. Uma vez que o acionamento do botão em si seria algo intrinsecamente desvalorizado, e não em virtude de suas consequências causais, o botão poderia ser algo completamente inerte, feito de massinha de modelar. Além disso, é irrelevante se a IA saberá ou não se o botão já havia sido pressionado. O essencial é que a IA acredite que o botão provavelmente permanecerá intocado se ela seguir agindo conforme o interesse do seu diretor em vez de se rebelar.

É possível que essa configuração seja refinada. Em vez de tentarmos dotar a IA com um objetivo final físico, como o exemplo do botão, poderíamos construir uma IA que tivesse como objetivo final receber uma série de “prêmios criptográficos”.[328] Esses prêmios consistiriam em sequências de números que serviriam como chaves criptográficas para cifras que teriam sido geradas antes da criação da IA e que teriam sido incorporadas ao seu sistema de motivação.[329] Essas sequências especiais de números seriam muito cobiçadas pela IA, constituindo-se em um tipo especial de prêmio que a IA não conseguiria obter por meio de *wireheading*.[330] As chaves poderiam ser armazenadas em um local seguro no qual pudessem ser rapidamente destruídas se a IA tentasse se apoderar delas. Caso a IA cooperasse, as chaves seriam distribuídas em um ritmo constante.

Poderíamos, então, dar à IA uma função de utilidade limitada e estabelecer um esquema no qual a IA receberia (por exemplo) 99% de sua utilidade máxima no primeiro prêmio; 99% de sua utilidade potencial remanescente no segundo prêmio; e assim por diante. Mantendo uma “taxa de desconto” alta e transmitindo os prêmios para a IA numa taxa razoavelmente alta, poderíamos dissuadir a IA de tentar executar planos de longo prazo que aumentassem o número esperado de prêmios num futuro remoto à custa de algum risco de que seu fluxo atual de prêmios pudesse ser diminuído, ou que o próximo prêmio pudesse ser cancelado ou

destruído. Se a IA sempre acreditar que desertar (ou seja, fazer qualquer outra coisa que não proteger os interesses do diretor da melhor maneira possível) corresponderia a um risco de pelo menos 2% de não receber seu próximo prêmio, enquanto a cooperação representaria uma redução de menos de 1%, então, uma IA que maximizasse a utilidade cooperaria sempre.

O que poderia ocorrer de errado em tal esquema de incentivo? Uma possibilidade é que a IA não venha a confiar que o operador humano entregará as recompensas prometidas. O histórico de confiabilidade humano é algo que difere bastante de uma linha reta, perfeita e infalível. A IA poderia ter receio de que o operador mudasse de ideia, aumentasse o nível de exigência ou não reconhecesse que a IA cumpriu com o seu papel. A IA também poderia se preocupar com a possibilidade de que o operador ficasse incapacitado e não conseguisse mais cumprir com o acordo. O risco combinado de tais falhas poderia exceder o risco da tentativa de tomada de controle do mecanismo de premiação. Até mesmo uma IA confinada que possuísse a panóplia de superpoderes já seria uma força poderosa. (No caso de uma IA que *não* estivesse confinada, subtrair o mecanismo de recompensa governado pelos humanos seria como tirar um doce da boca de uma criança.)

Outro problema desse esquema de incentivo é que ele pressupõe que nós teremos condições de saber se os resultados produzidos pela IA seriam ou não do nosso interesse. Como será visto nos próximos capítulos, essa pressuposição não é inócua.

Uma avaliação completa da viabilidade dos métodos de incentivo também teria que levar em conta uma série de outros fatores, incluindo algumas considerações esotéricas que poderiam talvez tornar tais métodos mais viáveis do que uma análise preliminar poderia sugerir. Particularmente, é possível que a IA enfrente uma incerteza indexical que seja impossível de ser eliminada caso ela não consiga ter certeza de que não está numa simulação de computador (em oposição a uma realidade física não simulada e “concreta”) e essa dúvida epistemológica pode influenciar radicalmente suas deliberações (ver quadro 8).

Inibição

Outro método possível de controle de capacidade seria limitar as faculdades intelectuais do sistema ou seu acesso à informação. Isso poderia ser feito se rodássemos a IA em um hardware lento ou com pouca memória. No caso de um sistema confinado, o fluxo de entrada de informações poderia ser restrito.

Inibir uma IA por esses meios limitaria sua utilidade. Esse método, portanto, enfrenta um dilema: se não houver inibição suficiente, a IA poderia ter a perspicácia de encontrar uma maneira para se tornar mais inteligente (e assim dominar o mundo); se inibirmos demasiadamente seu crescimento, a IA seria apenas mais um software banal qualquer. Uma IA radicalmente inibida certamente seria mais segura, mas isso não resolve a questão de como conseguir uma detonação controlada: uma explosão de inteligência continuaria sendo possível e seria provocada de forma simples por algum outro sistema qualquer, talvez apenas um pouco mais tarde.

Poderíamos pensar que seria seguro construir uma superinteligência, contanto que ela recebesse apenas dados limitados a fatos sobre alguma área específica de conhecimento. Por exemplo, seria possível construir uma IA que não tivesse

sensores e cuja memória estivesse pré-carregada apenas com fatos referentes à engenharia de petróleo ou à química de peptídios. Mas se essa IA for superinteligente — se ela tiver um nível super-humano de inteligência *geral* —, a privação de dados não seria uma garantia de segurança.

Há muitas razões para isso. Em primeiro lugar, a ideia de que uma informação seja apenas "a respeito" de certo tópico é, geralmente, problemática. Qualquer tipo de informação pode, em princípio, ser relevante para qualquer tipo de tópico, dependendo da bagagem anterior de informações de quem raciocina.[331] Além disso, um conjunto de dados contém informação não apenas sobre o domínio no qual os dados foram coletados, mas também a respeito de vários outros fatos circunstanciais. Uma mente astuta que tivesse acesso a uma base de conhecimento referente, nominalmente, apenas à química de peptídios poderia fazer inferências sobre uma vasta gama de outros tópicos. O fato de que certas informações estariam presentes e outras não poderia indicar para a IA o estágio em que se encontra a ciência humana, os métodos e instrumentos disponíveis para o estudo dos peptídios, as tecnologias necessárias para a fabricação desses instrumentos e a natureza dos cérebros e sociedades que conceberam tais estudos e instrumentos. Uma *superinteligência* seria capaz de fazer uma série de inferências corretas com base em algo que nossas ineptas mentes humanas consideram apenas migalhas de evidências. Até mesmo sem qualquer base de conhecimento determinada, uma mente suficientemente superior pode ser capaz de aprender muito simplesmente fazendo uma introspecção sobre o funcionamento de sua própria psique — as opções de projeto refletidas em seu código fonte e as características físicas de seus circuitos.[332] Talvez uma superinteligência pudesse até fazer muitas deduções à respeito das propriedades prováveis do mundo *a priori* (combinando inferência lógica com uma probabilidade *a priori* favorável à existência de mundos mais simples e alguns fatos elementares que podem ser derivados da existência da superinteligência enquanto um sistema racional). Ela poderia imaginar as consequências de diferentes leis da física possíveis: que tipos de planetas se formariam, que tipos de vida inteligente evoluiriam, que tipos de sociedade se desenvolveriam, que tipos de métodos para a resolução do problema de controle seriam testados e como poderiam ser derrotados.[333]

Por essas razões, seria uma imprudência confiar na privação de informação como o principal meio de controle do poder de uma superinteligência. Contudo, algumas formas de inibição, incluindo a limitação da capacidade de hardware e a prevenção do desenvolvimento de faculdades cognitivas específicas, poderiam ser úteis como métodos de controle auxiliares durante a fase de desenvolvimento de uma IA. Eles poderiam ser usados, por exemplo, para regular a taxa de ganho de capacidade em uma IA intermediária e para permitir que fossem feitos testes em vários estágios de desenvolvimento antes que a IA atingisse a plenitude de suas capacidades.

Detonadores

Os detonadores são mecanismos que realizam testes de diagnóstico num sistema (possivelmente sem seu conhecimento) e efetuam seu desligamento caso detectem sinais de perigo. Os programadores podem então examinar gravações das

atividades suspeitas quando quiserem e fazer mudanças nas bases de código ou no ambiente da IA antes de reiniciá-la. Se acharem que o problema é muito grave, eles poderiam decidir abortar todo o projeto.

Detonadores diferem dos métodos de incentivo na medida em que não dependem do conhecimento do sistema a respeito das consequências de se envolver em atividades proibidas. Detonadores estão mais estreitamente relacionados com os métodos de inibição. Assim como a inibição, detonadores poderiam ser usados como um sistema de defesa temporário, proporcionando um grau de proteção durante a fase de desenvolvimento da IA. Em princípio, detonadores também poderiam ser usados durante a fase operacional, especialmente em sistemas confinados. No entanto, a habilidade dos detonadores em conter uma superinteligência plena deve permanecer como algo muito duvidoso, uma vez que seria difícil se assegurar de que tal agente não seria capaz de encontrar maneiras de subverter qualquer detonador elaborado pelo intelecto humano.

Contudo, os detonadores são importantes como uma medida complementar de segurança durante a fase de desenvolvimento de uma IA. A codificação das expectativas em relação à maneira pela qual o sistema deveria supostamente funcionar seria uma prática importante para os criadores de uma IA, pois com base nisso eles poderiam cercar a IA de detonadores que seriam ativados caso qualquer uma dessas expectativas fosse violada. Para tornar os testes mais rigorosos, “iscas” poderiam ser estrategicamente colocadas de maneira que uma IA que não estivesse funcionando corretamente seria levada a cometer alguma violação que pudesse ser facilmente observável. Por exemplo, se uma IA tivesse sido projetada para não querer acessar a internet, uma porta falsa de ethernet poderia ser instalada como isca (que conduziria a um botão de desligamento automático) apenas para verificar se a IA tentaria usá-la. (Alguns exemplos de detonadores são descritos na tabela 9.)

Deve-se enfatizar que o valor de um detonador depende não só de seu próprio mecanismo, mas também — criticamente — de como um projeto reagiria quando esse detonador fosse acionado. Se os programadores ou gerentes do projeto, impacientes para obter algum progresso, simplesmente ligarem o sistema novamente depois que o detonador tivesse sido acionado — ou se ligarem o sistema após fazer modificações que evitassem que o detonador fosse acionado novamente —, então nenhuma segurança seria obtida, ainda que o detonador estivesse funcionando exatamente conforme o planejado.

TABELA 9

Diferentes tipos de detonadores

Tipo de monitoramento	Descrição e exemplos
Comportamento	Detectores poderiam ser colocados ao redor de uma IA confinada para detectar tentativas de fuga. Por exemplo, detectores poderiam interceptar tentativas de comunicação via rádio ou tentativas de acesso a recursos computacionais internos restritos. Uma porta de ethernet que funcionasse como uma espécie de “fruto proibido” poderia ser instalada: uma aparente conexão à internet que levaria ao botão de desligamento.

Habilidade	Testes de capacidade automatizados poderiam ser realizados em intervalos frequentes para determinar o nível de habilidade da IA em vários domínios. Se a taxa de melhoria for inesperadamente alta ou se a IA atingir um nível de competência que a deixasse próxima a uma zona potencialmente perigosa, o sistema poderia ser desacelerado ou pausado para permitir que os programadores monitorassem seu desenvolvimento mais de perto.
Conteúdo	Os processos internos da IA poderiam ser monitorados à procura de atividades suspeitas. Por exemplo, qualquer mudança na representação dos valores finais da IA poderia resultar em um desligamento automático para uma revisão. A maneira como a IA utiliza recursos e métodos algorítmicos poderia também ser monitorada contra padrões inesperados. De modo mais ambicioso, os pontos de vista da IA, planos, submetas e justificativas, poderiam ser continuamente inspecionados para assegurar que estejam de acordo com as expectativas dos programadores. A monitoração dos estados intencionais da IA possibilitaria uma forma particularmente potente de monitoramento de conteúdo: o escaneamento do processo cognitivo da IA como forma de encontrar a concepção de uma mentira — aquele momento vulnerável em que uma IA hostil elabora pela primeira vez a intenção de esconder suas verdadeiras intenções.* Monitoramentos de conteúdo que necessitem que os estados intencionais da IA sejam transparentes para os programadores ou para um mecanismo de monitoramento automático podem talvez não ser factíveis para todos os tipos de arquiteturas de IA. (Algumas redes neurais, por exemplo, são opacas, uma vez que representam informações de maneira holística e de formas que não necessariamente têm relação com conceitos humanos.) Essa talvez pode ser uma razão para evitar o uso de tais arquiteturas.

* No momento *seguinte* à concepção de uma mentira, a IA poderia dar um jeito de apagar os traços de seu pensamento rebelde. Seria importante, portanto, que esse detonador operasse continuamente. Outra boa alternativa seria usar uma “caixa-preta” que armazenasse informações completas a respeito de todas as atividades da IA (incluindo o tempo exato de digitação de todos os programadores), de modo que sua trajetória pudesse ser traçada novamente ou analisada após um desligamento automático. Essas informações poderiam ser armazenadas em discos óticos não regraváveis.

Métodos de seleção de motivação

Os métodos de seleção de motivação buscam prevenir resultados indesejáveis por meio de uma predefinição do que a superinteligência terá vontade de fazer. Ao projetar o sistema de motivação do agente e seus objetivos finais, esses métodos produziriam uma superinteligência que *não desejaria* se aproveitar de uma vantagem estratégica decisiva de forma nociva. Uma vez que um agente superinteligente será habilidoso em alcançar seus propósitos, caso ele prefira não causar dano (“dano” sendo pensado aqui em algum sentido relevante do termo), então ele teria a tendência de não causar dano (novamente num sentido relevante de “dano”).

A seleção de motivação pode envolver a formulação explícita de um objetivo ou de um conjunto de regras que deverão ser seguidas (*especificação direta*) ou a configuração do sistema de modo que ele possa descobrir um conjunto apropriado de valores por si mesmo, utilizando-se de critérios formulados implicitamente ou indiretamente (*normatividade indireta*). Uma opção em relação à seleção de motivação seria tentar construir o sistema de modo que ele tivesse objetivos modestos e pouco ambiciosos (*domesticidade*). Uma alternativa à criação de um sistema de motivação partindo do zero poderia se dar por meio da seleção de um

agente que já tenha um sistema de motivação aceitável e, a partir daí, aumentar seus poderes cognitivos para transformá-lo em uma superinteligência, assegurando-nos de que o sistema de motivação não se torne corrupto durante o processo (*ampliação*). Analisaremos cada uma dessas opções individualmente.

Especificação direta

A especificação direta é a abordagem mais objetiva para o problema do controle. Essa abordagem possui duas versões possíveis, uma baseada em regras e outra consequencialista, e envolve a tentativa de definir explicitamente um conjunto de regras ou valores que fariam com que até mesmo uma IA superinteligente livre agisse de forma segura e benéfica. A especificação direta, entretanto, encontra alguns obstáculos que podem se tornar insuperáveis, derivados tanto das dificuldades em determinar quais regras ou valores nós desejaríamos que guiassem essa IA como das dificuldades em expressar essas regras ou valores em programas de computador.

O exemplo tradicional da abordagem direta baseada em regras é o conceito das "três leis da robótica", formuladas pelo autor de ficção científica Isaac Asimov em um conto publicado em 1942.[334] As três leis são: (1) Um robô não pode ferir um ser humano ou, por inação, permitir que um ser humano sofra algum mal; (2) Um robô deve obedecer às ordens que lhe sejam dadas por seres humanos, exceto nos casos em que tais ordens entrem em conflito com a Primeira Lei; (3) Um robô deve proteger sua própria existência desde que tal proteção não entre em conflito com a Primeira ou Segunda Leis. É constrangedor para a nossa espécie que as leis de Asimov tenham permanecido, por mais de meio século, como o que havia de mais avançado nesse tema: mesmo apesar dos problemas óbvios que essa abordagem apresenta, e alguns deles, inclusive, foram explorados nos próprios textos de Asimov (o autor provavelmente formulou essas leis precisamente para que elas falhassem de formas interessantes, fornecendo complicações de enredo férteis para suas histórias).[335]

Bertrand Russell, que passou muitos anos trabalhando nos fundamentos da matemática, observou certa vez que "tudo é vago em um grau imperceptível até que você tente tornar preciso".[336] A máxima de Russell se encaixa perfeitamente à abordagem da especificação direta. Considere, por exemplo, como poderíamos explicar em detalhes a primeira lei de Asimov. Ela quer dizer que um robô deveria minimizar a probabilidade de que qualquer ser humano pudesse sofrer algum mal? Nesse caso, as outras leis se tornariam supérfluas, uma vez que sempre é possível para a IA tomar alguma medida que tenha pelo menos algum efeito microscópico na probabilidade de um ser humano sofrer algum mal. Como o robô poderia comparar um risco grande de que alguns poucos humanos sofram algum mal com um risco pequeno de que muitos humanos sofram algum mal? E qual seria mesmo a definição precisa de "mal"? Como o "mal" da dor física deveria ser comparado ao "mal" da feiura arquitetônica ou da injustiça social? Um sádico seria prejudicado se ele fosse impedido de atormentar sua vítima? Como definimos "ser humano"? Por que nenhuma consideração é dada a outros seres consideráveis moralmente, tais

como animais conscientes não humanos e mentes digitais? Quanto mais ponderamos, mais as questões se proliferam.

Talvez o análogo mais próximo a um conjunto de regras que poderiam governar as ações de uma superinteligência operando livremente no mundo seria o sistema legal. Mas os sistemas legais foram desenvolvidos por meio de um longo processo de tentativa e erro, e regulam sociedades humanas que mudam em velocidades relativamente lentas. As leis podem ser revistas quando necessário. E, ainda mais importante, os sistemas legais são administrados por juízes e júris que geralmente utilizam bom senso e decência para ignorar interpretações legais que sejam logicamente possíveis, embora obviamente imprevistas e indesejadas pelos legisladores. É provável que seja humanamente impossível formular explicitamente um conjunto de regras altamente complexas de forma detalhada, aplicá-las em diversas circunstâncias e ainda conseguir fazer isso da maneira correta na primeira tentativa de implementação.[337]

Os problemas referentes à abordagem direta consequencialista são similares àqueles da abordagem direta baseada em regras. Isso é verdade inclusive se a IA tiver sido planejada para servir algum propósito aparentemente simples, tal como implementar uma versão do utilitarismo clássico. Por exemplo, o objetivo "Maximize o equilíbrio esperado entre o prazer e a dor presentes no mundo" pode parecer simples. Contudo expressá-lo em um programa de computador envolveria, entre outras coisas, especificar como reconhecer prazer e dor. Fazer isso de forma confiável pode demandar a resolução de uma série de problemas persistentes de filosofia da mente — apenas para obter uma descrição correta em linguagem natural. Então, essa descrição precisaria ser, de algum modo, traduzida para uma linguagem de programação.

Um pequeno erro tanto na representação filosófica como em sua transposição para um programa de computador poderia ter consequências catastróficas. Considere uma IA que tenha o hedonismo como seu objetivo final e que, portanto, queira preencher o universo com "hedônio" (matéria organizada em uma configuração ideal para a criação de experiências prazerosas). Para tanto, essa IA poderia produzir computrônio (matéria organizada em uma configuração ideal para realizar computações) e usá-lo para implementar mentes digitais em estado de euforia. Para apresentar uma eficiência máxima, a IA omitiria dessa implementação quaisquer faculdades mentais que não fossem essenciais para a experiência do prazer e exploraria quaisquer atalhos computacionais que, de acordo com a sua definição de prazer, não prejudicassem a criação de prazer. Por exemplo, essa IA poderia limitar sua simulação a circuitos de recompensa, omitindo faculdades tais como memória, percepção sensorial, função executiva e linguagem; ela poderia simular mentes em um nível relativamente grosseiro de funcionalidade, omitindo processos neuronais de baixo nível; ela poderia substituir computações comumente repetidas por chamadas que buscam valores em uma tabela de pesquisa pré-definida; ou ela poderia pôr em prática algum esquema em que várias mentes compartilhassem a maior parte dos seus maquinários computacionais subjacentes (suas "bases de superveniência" no linguajar filosófico). Esses truques poderiam aumentar consideravelmente a quantidade de prazer passível de ser produzido com

uma dada quantidade de recursos. Não é claro o quanto isso seria desejável. Além do mais, caso esteja equivocado o critério da IA para determinar se um processo físico gera prazer, então as otimizações da IA poderiam colocar tudo a perder: ela descartaria algo que não é essencial de acordo com seus critérios, mas que poderia ser essencial de acordo com o critério implícito em nossos valores humanos. O universo, então, se preencheria não só com hedônio fervilhante de alegria, mas também com processos computacionais inconscientes que seriam completamente inúteis — o equivalente a adesivos de rostos felizes fotocopiados trilhões e trilhões de vezes e colados por todo o universo.

Domesticidade

A autolimitação é um tipo especial de objetivo final que pode ser mais favorável à especificação direta do que os exemplos fornecidos anteriormente. Enquanto parece ser extremamente difícil especificar como alguém gostaria que uma superinteligência viesse a se comportar no mundo de uma maneira *geral* — uma vez que isso nos obrigaria a levar em conta todas as comparações entre todas as situações que pudessem vir a ocorrer —, talvez fosse viável especificar como uma superinteligência deveria se comportar em uma situação particular. Desse modo, nós poderíamos tentar motivar o sistema a restringir suas ações a uma escala pequena, dentro de um contexto reduzido e utilizando apenas um conjunto limitado de modos de ação. Vamos chamar essa abordagem, que consiste em dotar a IA com objetivos finais cuja intenção é limitar a abrangência de suas ambições e atuação, de "domesticidade".

Por exemplo, poderíamos tentar projetar uma IA de maneira que ela funcionaria apenas como um dispositivo que responda a perguntas (um "oráculo", para antecipar a terminologia que apresentaremos no próximo capítulo). Simplesmente dotar a IA com o objetivo final de responder qualquer pergunta feita com a máxima precisão poderia ser pouco seguro — vale lembrar a "catástrofe da hipótese de Riemann" descrita no capítulo 8. (Devemos refletir também que esse objetivo incentivaria a IA a tomar medidas para garantir que lhe sejam feitas perguntas fáceis.) Para atingir a domesticidade, poderíamos tentar definir um objetivo final que de algum modo superasse essas dificuldades: talvez um objetivo que combinasse as aspirações de responder perguntas corretamente e de minimizar o impacto da IA no mundo, exceto quando esse impacto fosse resultado de uma consequência acidental do fornecimento de respostas precisas e não manipuladoras às perguntas que fossem feitas.[338]

A especificação direta de tal objetivo de domesticidade é, provavelmente, mais viável do que a especificação direta de um objetivo mais ambicioso ou de um conjunto completo de regras de operação em uma série de situações indefinidas *a priori*. Mesmo assim, desafios significativos continuariam a existir. Teríamos que tomar cuidado, por exemplo, com a definição do que significaria para a IA a ideia de "minimizar seu impacto no mundo", de modo a assegurar que a maneira como a IA mediria impacto coincidiria com os nossos padrões do que seria um impacto grande ou pequeno. Uma definição ruim conduziria a um resultado ruim. Há também outros

tipos de riscos associados à construção de um oráculo, que discutiremos mais adiante.

Há uma compatibilidade natural entre a abordagem de domesticidade e o confinamento físico. Devemos tentar "confinar" uma IA de tal maneira que ela seja *incapaz* de escapar enquanto tentamos simultaneamente moldar seu sistema de motivação de forma que ela *não deseje* escapar, ainda que encontre meios para tanto. Se tudo mais se mantiver constante, a existência de múltiplos mecanismos de segurança independentes deveria aumentar as chances de sucesso.[339]

Normatividade indireta

Se a especificação direta parecer muito difícil, poderíamos tentar usar a normatividade indireta em seu lugar. A ideia básica aqui é que, em vez de especificar diretamente um padrão normativo concreto, nós especificaríamos um processo para estabelecer um padrão. Poderíamos, então, construir o sistema de modo que ele fosse motivado a executar esse processo e adotar qualquer padrão que derive do processo executado.[340] Por exemplo, o processo poderia ser a condução de uma investigação a respeito da questão empírica do que uma versão humana idealizada preferiria que a IA fizesse. O objetivo final fornecido à IA nesse exemplo poderia ser algo como: "Realize aquilo que nós teríamos desejado que você realizasse se nós tivéssemos pensado sobre o assunto intensamente e durante muito tempo".

Explicações mais detalhadas sobre a normatividade indireta serão dadas no capítulo 13. É nele que revisitaremos a ideia da "extrapolação da nossa vontade" e exploraremos várias formulações alternativas. A normatividade indireta é uma abordagem muito importante para a seleção de motivação. O seu potencial reside no fato de que ela permitiria que deixássemos a cargo da superinteligência a maior parte do trabalho cognitivo mais difícil exigido para a elaboração da especificação direta de um objetivo final apropriado.

Ampliação

O último método de seleção de motivação na nossa lista é a ampliação. A ideia aqui é iniciar com um sistema que já tivesse um sistema motivacional aceitável e melhorar suas faculdades cognitivas para torná-lo superinteligente em vez de tentar projetar um sistema de motivação do zero. Se tudo corresse bem, isso nos daria uma superinteligência com um sistema motivacional satisfatório.

Essa abordagem, obviamente, seria inútil no caso de uma IA embrionária recém-criada. Mas a ampliação é um método potencial de seleção de motivação para outros caminhos que levam até uma superinteligência, incluindo a emulação completa do cérebro, melhoramento biológico, interfaces cérebro-computador, além de redes e organizações, onde existe a possibilidade de construir o sistema a partir de um núcleo normativo (seres humanos normais) que já tenha uma representação de valores humanos.

A atratividade da ampliação pode aumentar em proporção à nossa falta de esperança em relação às outras abordagens para o problema do controle. Criar um sistema de motivação para uma IA embrionária que pudesse permanecer seguro e

benéfico, de forma confiável, enquanto estivesse ocorrendo um autoaperfeiçoamento recursivo e mesmo conforme o sistema se tornasse uma superinteligência madura, seria uma tarefa bem difícil, especialmente se tivermos que chegar a essa solução já na primeira tentativa. Com a ampliação, pelo menos teríamos a oportunidade de iniciar o processo com um sistema que tivesse motivações familiares e semelhantes às humanas.

Olhando pelo lado negativo, seria difícil assegurar que um sistema de motivação complexo, evoluído, desengonçado e ainda mal entendido como o dos seres humanos não fosse se corromper quando seu mecanismo cognitivo explodisse em direção à estratosfera. Como já foi discutido, um procedimento de emulação do cérebro imperfeito que preservasse o funcionamento intelectual poderia não conseguir preservar todos os aspectos da personalidade do cérebro emulado. O mesmo valeria (embora talvez em menor grau) para aperfeiçoamentos biológicos da cognição, que poderiam afetar sutilmente a motivação, e valeria também para melhorias na inteligência coletiva de organizações e redes que poderiam mudar, negativamente, suas dinâmicas sociais (por exemplo, poderia ocorrer uma degradação das atitudes do coletivo em relação aos que se encontrassem fora da organização ou em relação a alguns de seus próprios constituintes). Se a superinteligência for alcançada por quaisquer um desses meios, um patrocinador do projeto teria dificuldade em garantir quais seriam as motivações principais desse sistema quando ele chegasse à maturidade. Uma IA com uma arquitetura matematicamente bem especificada e com fundamentos elegantes poderia — apesar de toda a sua diversidade não antropomórfica — oferecer uma transparência maior, e talvez existisse até mesmo a possibilidade de que aspectos importantes da sua funcionalidade fossem formalmente verificados.

No fim das contas, entretanto, se compararmos as vantagens e desvantagens da ampliação, a opção de confiar ou não nesse processo pode estar além de nós. Se a superinteligência for primeiramente alcançada pelo caminho da inteligência artificial, a ampliação não é aplicável. Se, ao contrário, a superinteligência for primeiramente alcançada por um algum caminho que não envolva a inteligência artificial, então muitos dos outros métodos de seleção de motivação talvez não possam ser aplicados. Ainda assim, pontos de vista a respeito da probabilidade de sucesso da ampliação apresentam relevância estratégica, contanto que tenhamos oportunidade de influir no tipo de tecnologia que será responsável pela produção da primeira superinteligência.

TABELA 10

Métodos de controle

Controle de capacidade	
Métodos de confinamento	O sistema é confinado de modo que apenas é capaz de afetar o mundo externo por meio de um canal restrito e pré-aprovado. Inclui métodos de confinamento físico e restrição de acesso à informação.

Métodos de incentivo	O sistema é colocado dentro de um ambiente que fornece incentivos apropriados. Isso poderia envolver integração social em um mundo com entidades igualmente poderosas. Outra variação seria o uso de prêmios criptográficos. “Captura antrópica” também é uma possibilidade muito relevante, porém envolveria considerações esotéricas.
Inibição	São impostas restrições às capacidades cognitivas do sistema ou à sua habilidade em afetar processos internos centrais.
Detonadores	Testes de diagnóstico são realizados no sistema (possivelmente sem o seu conhecimento) e um mecanismo automático desliga o mesmo se alguma atividade perigosa for detectada.
Seleção de motivação	
Especificação direta	O sistema é dotado com algum sistema de motivação diretamente especificado, que pode ser consequencialista ou envolver o de um conjunto de regras específicas.
Domesticidade	Um sistema de motivação é projetado para limitar severamente o escopo das atividades e ambições do agente.
Normatividade indireta	A normatividade indireta poderia envolver princípios consequencialistas ou baseados em regras, mas se distingue por ter como base a confiança em uma abordagem indireta para a especificação das regras que devem ser seguidas ou dos valores que serão alcançados.
Ampliação	Na ampliação, iniciamos com um sistema que já tenha motivações substancialmente humanas ou benevolentes e aperfeiçoamos suas capacidades cognitivas para torná-lo superinteligente.

Sinopse

Faremos aqui uma breve sinopse antes de encerrar este capítulo. Começamos diferenciando duas classes bem amplas de métodos para lidar com o problema de agência que são centrais para as questões de segurança em uma IA: controle de capacidade e seleção de motivação. A tabela 10 resume essas classes.

Cada método de controle possui vulnerabilidades em potencial e apresenta diferentes graus de dificuldade para sua implementação. Podemos pensar, talvez, que deveríamos classificá-los do melhor para o pior e, então, optar pelo melhor método. Mas isso seria simplório. Alguns métodos podem ser usados de forma combinada, enquanto outros são exclusivos. Mesmo um método comparativamente inseguro pode ser aconselhável se puder ser facilmente utilizado como um método adicional, ao passo que um método forte pode não ser atraente caso limite o uso de outros dispositivos de segurança desejáveis.

Desse modo, é necessário considerar quais combinações fazem sentido. Precisamos levar em consideração que tipos de sistema tentaremos construir e

quais métodos de controle seriam aplicáveis a cada um deles. Esse será o tópico tratado no próximo capítulo.

10

CAPÍTULO 10

ORÁCULOS, GÊNIOS, SOBERANOS E FERRAMENTAS

Alguns poderiam dizer: “Construa apenas um sistema que responda a perguntas!” ou “Construa apenas uma IA que seja uma ferramenta, e não um agente!”. Mas essas sugestões não fazem com que todas as preocupações relativas à segurança desapareçam e não é uma questão trivial definir que tipo de sistema ofereceria as melhores perspectivas de segurança. Consideraremos quatro tipos ou “castas” — oráculos, gênios, soberanos e ferramentas — e explicaremos como cada uma se relaciona com as outras.[341] Todas oferecem um conjunto diferente de vantagens e desvantagens em nossa busca pela solução do problema de controle.

Oráculos

Um oráculo é um sistema que responde a perguntas. Ele pode aceitar perguntas em uma linguagem natural e apresentar respostas em forma de texto. Um oráculo que aceite apenas perguntas do tipo sim ou não poderia apresentar seu melhor palpite na forma de um único bit, ou talvez usando alguns bits extras, que representassem o seu grau de confiança na resposta. Um oráculo que aceite perguntas abertas precisaria de algum tipo de métrica com a qual classificaria possíveis respostas verdadeiras em termos de seu valor informativo e pertinência.[342] Em ambos os casos, construir um oráculo que tenha uma habilidade de domínio geral para responder a perguntas em linguagem natural é um problema IA-completo. Se isso pudesse ser feito, é provável que fosse também possível construir uma IA que tivesse uma habilidade aceitável para compreender não apenas palavras, mas também intenções humanas.

Oráculos com formas de superinteligência de domínio restrito também são concebíveis. Por exemplo, é possível imaginar um oráculo de matemática que aceitasse apenas questões representadas em uma linguagem formal, mas que fosse muito eficiente em respondê-las (por exemplo, ele seria capaz de resolver em um instante praticamente qualquer problema matemático formalmente representado, cuja solução apenas poderia ser obtida por matemáticos humanos por meio de um trabalho colaborativo que duraria séculos). Esse oráculo de matemática funcionaria como uma espécie de primeiro degrau em direção a uma superinteligência de domínio geral.

Oráculos com superinteligência restrita a domínios extremamente limitados já existem. Uma calculadora de bolso pode ser considerada como um oráculo bastante restrito para questões de aritmética básica; uma ferramenta de busca na internet pode ser vista como uma realização bastante parcial de um oráculo cujo domínio engloba uma parte significativa do conhecimento humano declarativo. Esses oráculos de domínio limitado são ferramentas, e não agentes (mais adiante,

veremos mais a respeito das IAs ferramentas). No que se segue, entretanto, o termo “oráculo” fará referência a sistemas superinteligentes de domínio geral que respondam a perguntas, salvo indicação em contrário.

Para fazer com que uma superinteligência geral funcione como um oráculo, poderíamos aplicar seleção de motivação e controle de capacidade. Seleção de motivação talvez seja algo mais simples de ser feito num oráculo do que em outras castas de superinteligência, pois o objetivo final de um oráculo poderia ser mais simples. Gostaríamos que o oráculo fornecesse respostas honestas, e não manipuladoras, e que limitasse seu impacto no mundo. Aplicando um método de domesticidade, poderíamos exigir que o oráculo usasse apenas recursos designados para produzir suas respostas. Por exemplo, nós poderíamos estipular que ele baseasse suas respostas em um conjunto predefinido de informações, tal como uma versão armazenada de todo o conteúdo da internet em um dado momento, sendo que não lhe seria permitido usar mais do que um número pré-fixado de etapas computacionais.[343] Para não incentivar o oráculo a nos manipular de maneira que façamos apenas perguntas simples — o que aconteceria se lhe fosse fornecido o objetivo de maximizar sua exatidão para todas as perguntas realizadas —, poderíamos fornecer apenas o objetivo de responder a uma única pergunta e se desligar logo depois que a resposta fosse dada. A pergunta seria pré-carregada em sua memória antes que o programa fosse rodado. Para fazer uma segunda pergunta, teríamos que reiniciar a máquina e rodar o mesmo programa com uma pergunta diferente pré-carregada na memória.

Desafios sutis e potencialmente traiçoeiros surgem mesmo na especificação de um sistema de motivação relativamente simples que é necessário para guiar um oráculo. Suponhamos, por exemplo, que elaboremos alguma explicação a respeito do que significa para a IA “minimizar seus impactos no mundo, subordinando-se ao alcance de certos resultados” ou “usar apenas recursos predefinidos na preparação das respostas”. O que aconteceria se a IA, ao longo de seu desenvolvimento intelectual, passasse pelo equivalente a uma revolução científica que envolvesse uma mudança em sua ontologia básica?[344] Inicialmente, os termos “impacto” e “recursos designados” seriam explicados por nossa própria ontologia (postulando a existência de vários objetos físicos, tais como computadores). Mas da mesma forma que abandonamos as categorias ontológicas que eram consideradas naturais por cientistas em outras épocas (“flogisto”, “elã vital” e “simultaneidade absoluta”), uma IA superinteligente poderia descobrir que algumas de nossas categorias existentes são baseadas em concepções fundamentalmente equivocadas. O sistema de objetivos de uma IA que estivesse passando por uma crise ontológica precisaria ser resiliente o suficiente para que o “espírito” original do conteúdo de seus objetivos fosse mantido e caridosamente transposto para sua nova ontologia.

Mesmo considerando que tornar um oráculo seguro através da seleção de motivação não seja nada trivial, talvez esse processo seja mais fácil do que fazer o mesmo com uma IA que esteja vagando pelo mundo em busca de algum objetivo complicado. Esse argumento defende que a primeira superinteligência deveria ser preferencialmente um oráculo. Outro ponto a favor de um caminho que siga na direção de um oráculo diz respeito à sua maior suscetibilidade ao controle de

capacidade. Todos os métodos padrão de confinamento se encaixam aqui. Também podem existir métodos especificamente aplicáveis a um oráculo. Considere, por exemplo, o risco de que um oráculo forneça respostas que não sejam totalmente verdadeiras, mas que sutilmente nos manipulem para a realização de seus próprios objetivos secretos. Uma maneira de diminuir um pouco essa ameaça seria criando múltiplos oráculos, cada um programado de maneira ligeiramente diferente dos demais e com bases de informação também ligeiramente distintas entre si. Um mecanismo simples poderia então comparar todas as respostas fornecidas pelos diferentes oráculos e apenas apresentá-las aos humanos caso todas as respostas fossem semelhantes. Entretanto, devemos considerar que qualquer oráculo superinteligente de domínio geral estaria ciente da existência de outros oráculos (ele poderia inferir isso a partir de seu conhecimento sobre a sociedade humana), mas talvez fosse possível esconder detalhes de implementação de cada um dos oráculos, evitando que esses detalhes fossem vistos pelos outros. Se os oráculos não puderem conversar entre si, então talvez fosse difícil coordenar uma maneira de responder às perguntas de forma manipuladora. Existem muitas maneiras de se desviar da verdade e os oráculos poderiam discordar entre si a respeito de qual seria a mais atrativa — enquanto a verdade em si é um ponto de Schelling (um lugar relevante para a concordância na ausência de comunicação). Então, se os oráculos chegam a um consenso, isso poderia ser um sinal de que a resposta é verdadeira.[345]

Um oráculo seria idealmente confiável se pudéssemos assumir seguramente que ele responde sempre da maneira mais precisa possível de acordo com suas habilidades. Mas mesmo um oráculo pouco confiável poderia ser útil. Nós poderíamos fazer ao oráculo perguntas com respostas difíceis de ser encontradas, mas cuja verificação seja mais fácil. Muitos problemas matemáticos têm essa característica. Caso estivermos ponderando se uma dada proposição matemática é verdadeira, poderíamos pedir que o oráculo produzisse uma prova ou uma refutação para tal proposição. Encontrar a prova pode exigir insight e criatividade que estão além de nossas capacidades, mas é possível conferir a validade de uma suposta prova através de um simples processo mecânico.

Se for dispendioso verificar as respostas (como geralmente ocorre em tópicos que não dizem respeito à lógica e à matemática), poderíamos selecionar aleatoriamente um subconjunto de respostas dadas pelo oráculo para verificação. Se todas estiverem corretas, podemos atribuir uma probabilidade alta de que a maioria das outras respostas também estarão. Esse truque permitiria uma grande economia com a verificação de respostas confiáveis, algo que seria muito custoso realizar individualmente. (Infelizmente, isso não poderia garantir respostas confiáveis para perguntas que *não saibamos* verificar, uma vez que um oráculo dissimulado poderia optar por responder corretamente apenas às perguntas cujas respostas, segundo ele, nós seríamos capazes de verificar.)

É possível que existam algumas maneiras pelas quais poderíamos nos beneficiar de uma indicação de resposta correta (ou de um direcionamento a um método para encontrar a resposta correta), ainda que sejamos obrigados a desconfiar ativamente de sua procedência. Por exemplo, poderíamos perguntar pela solução de vários

problemas técnicos ou filosóficos que surgiriam ao longo do desenvolvimento de métodos mais avançados de motivação de seleção. Caso desenvolvêssemos uma IA que acreditássemos ser segura, poderíamos pedir ao oráculo que identificasse alguma falha significativa no projeto e solicitar que ele nos explicasse essa falha em vinte palavras ou menos. Com perguntas dessa natureza, poderíamos obter informações valiosas. Entretanto, muito cuidado e moderação seriam necessários para que nós não fizéssemos perguntas *demais* — e não nos permitíssemos dar *muitos* detalhes a respeito das respostas fornecidas —, pois quanto menos detalhes um oráculo pouco confiável tivesse, menor seria a chance de que ele nos influenciasse psicologicamente (através de respostas aparentemente plausíveis, mas sutilmente manipuladoras). Não seriam necessários muitos bits de comunicação para que uma IA com o superpoder de manipulação social nos dobrasse às suas vontades.

Ainda que o oráculo em si funcione exatamente como o planejado, existe o risco de que ele seja mal utilizado. Uma dimensão óbvia desse problema é que uma IA oráculo poderia ser uma fonte de imenso poder, que daria uma vantagem estratégica decisiva ao seu operador. Esse poder poderia ser ilegítimo e não utilizado para o bem comum. Outra possibilidade mais sutil, porém não menos importante, é que o uso de um oráculo poderia representar um grande perigo para o próprio operador. Preocupações similares (que envolvem questões tanto filosóficas quanto técnicas) também surgem em relação às outras castas de superinteligência. Iremos explorá-las mais a fundo no capítulo 13. É suficiente notar aqui a grande importância de um protocolo que determine quais perguntas serão feitas, em que sequência e como as respostas serão apresentadas e disseminadas. Também é possível considerar a possibilidade de construir um oráculo de forma que ele se recusasse a responder a uma pergunta caso previsse que sua resposta teria consequências consideradas catastróficas de acordo com alguns critérios aproximados.

Gênios e soberanos

Um gênio é um sistema de execução de comandos: ele recebe um comando, o executa e então fica parado à espera do próximo comando.[346] Um soberano é um sistema que tem um mandato aberto para operar no mundo em busca de objetivos amplos e possivelmente de prazo muito longo. Embora ambos pareçam modelos que divergem radicalmente do que uma superinteligência poderia ser ou fazer, essa diferença não é tão profunda como pode parecer à primeira vista.

Com um gênio, já abrimos mão da propriedade mais atrativa de um oráculo: a oportunidade de utilizarmos métodos de confinamento. Podemos até considerar a criação de um gênio fisicamente confinado, por exemplo, um que consiga construir objetos apenas dentro de um volume designado — volume este que pode ficar isolado por paredes maciças ou por uma barreira carregada de explosivos armados para detonar caso o confinamento seja violado —, mas seria difícil confiar muito na segurança de qualquer método de confinamento físico para conter uma superinteligência equipada com materiais de construção e manipuladores versáteis.

Mesmo que fosse possível garantir um ambiente tão seguro quanto aquele que poderia ser obtido para um oráculo, não está claro quais seriam os benefícios de fornecer a uma superinteligência o acesso direto a manipuladores, em vez de fazê-la produzir um plano que pudéssemos inspecionar e depois usá-lo para alcançar os mesmos resultados. O ganho em velocidade e conveniência em se evitar a intermediação humana não parece compensar a perda dos já referidos usos dos métodos de confinamento mais fortes disponíveis para a contenção de um oráculo.

Se *fôssemos* criar um gênio, seria preferível construí-lo de modo que ele obedecesse às intenções por trás do comando em vez de seu significado literal, uma vez que um gênio que entendesse tudo literalmente (e que fosse superinteligente o suficiente para obter uma vantagem estratégica decisiva) poderia ser propenso a matar seu usuário e o resto da humanidade na primeira vez que fosse utilizado em virtude das razões explicadas na seção sobre modos de falha malignos, no capítulo 8. De forma mais abrangente, parece importante que um gênio busque uma forma caridosa — que poderia ser considerada razoável pelos seres humanos — de interpretar os comandos e que ele fosse motivado a executar cada comando a partir dessa linha de interpretação, em vez de fazê-lo de forma literal. O gênio ideal seria semelhante a um "supermordomo", e não a um "sábio autista".

No entanto, um gênio que possuísse as características de um supermordomo não estaria longe de se tornar um membro da casta dos soberanos. Considere, a título de comparação, a ideia de construir um soberano cujo objetivo final fosse o de obedecer ao espírito dos comandos que teríamos dado, caso tivéssemos construído um gênio em vez de um soberano. Tal soberano imitaria um gênio. Por ser superinteligente, esse soberano seria muito bem-sucedido em adivinhar quais comandos nós teríamos dado a um gênio (e ele poderia sempre nos perguntar a respeito, se isso pudesse ajudar na sua tomada de decisão). Existiria então alguma diferença realmente importante entre um soberano desse tipo e um gênio? Ou, buscando a distinção a partir do outro lado, considere que um gênio superinteligente pudesse, da mesma maneira, ser capaz de prever quais comandos lhe seriam fornecidos: o que ganharíamos se ele tivesse que esperar pelo comando para poder agir?

Podemos pensar que a grande vantagem de um gênio em relação a um soberano é que, se algo der errado, nós poderíamos fornecer um novo comando para o gênio para que ele parasse ou revertesse os efeitos das ações anteriores, enquanto um soberano simplesmente seguiria seu curso de ação, independentemente de nossos protestos. Mas essa aparente vantagem de segurança de um gênio é, em grande medida, ilusória. Os botões de "pare" ou "reverta" em um gênio funcionam apenas para modos de falha benignos: no caso de uma falha maligna — na qual, por exemplo, continuar com o comando dado tenha se tornado o objetivo final do gênio —, ele simplesmente desconsideraria qualquer tentativa subsequente de revogar o comando anterior.[347]

Uma opção seria tentar construir um gênio de modo que ele apresentasse automaticamente ao usuário uma previsão dos aspectos mais evidentes a respeito dos prováveis resultados do comando proposto e pedisse uma confirmação antes

de proceder sua execução. Tal sistema poderia ser denominado de *gênio com pré-visualização*. Mas se fosse possível fazer isso com um gênio, seria igualmente possível fazê-lo com um soberano. De modo que, novamente, esse não é um bom diferenciador entre um gênio e um soberano. (Supondo que uma funcionalidade de pré-visualização pudesse ser criada, as questões a respeito do seu uso são bem menos óbvias do que poderia parecer, a independer do grande apelo que pode haver na possibilidade de se obter uma pequena prévia dos resultados antes que nos arrisquemos a torná-los uma realidade irrevogável. Retornaremos a essas questões mais à frente.)

A habilidade de uma casta em imitar a outra se estende também aos oráculos. Um gênio poderia agir como um oráculo, caso responder apenas a alguns tipos de perguntas fosse o único comando fornecido a ele. Um oráculo, por sua vez, poderia substituir um gênio se lhe perguntássemos qual é o modo mais fácil de conseguir que certos comandos sejam executados. O oráculo seria capaz de nos fornecer instruções passo a passo para que alcançássemos o mesmo resultado que um gênio produziria, ou ele poderia até nos fornecer o código fonte de um gênio.[348] Argumentos similares podem ser feitos sobre a relação entre um oráculo e um soberano.

A verdadeira diferença entre as três castas não reside, então, nas derradeiras capacidades que elas poderiam vir a ter. Ao contrário, a diferença se resume às diferentes abordagens possíveis para o problema do controle. Cada casta corresponde a um conjunto diferente de medidas de precaução. A característica mais proeminente de um oráculo é que ele pode ser confinado. Também seria possível aplicar a domesticidade como seleção de motivação para um oráculo. É mais difícil confinar um gênio, mas pelo menos a domesticidade poderia ser aplicável. Um soberano não pode ser nem confinado nem manejado através da abordagem de domesticidade.

Se esses fossem os únicos fatores relevantes, a ordem de preferência seria clara: um oráculo seria mais seguro que um gênio, e este por sua vez seria mais seguro que um soberano; e quaisquer diferenças iniciais em conveniência e velocidade de operação seriam relativamente pequenas e facilmente ultrapassadas pelos ganhos em segurança que podem ser obtidos com a construção de um oráculo. Existem, entretanto, outros fatores que precisam ser levados em consideração. Quando escolhemos entre as castas, devemos considerar não apenas os riscos apresentados pelo sistema em si, mas também os perigos causados pelas formas como ele poderia ser utilizado. Obviamente, um gênio disponibiliza ao seu controlador um poder enorme, mas o mesmo vale para um oráculo.[349] Um soberano, ao contrário, poderia ser construído de maneira que não concederia a nenhuma pessoa ou grupo qualquer influência especial em relação aos seus resultados, e de forma que pudesse resistir a qualquer tentativa de corromper ou alterar seus objetivos originais. E mais: se a motivação de um soberano for definida por meio da “normatividade indireta” (um conceito que será descrito no capítulo 13), então ele poderia ser usado para alcançar algum resultado abstratamente definido, como “o que quer que seja mais justo e moralmente correto” — sem que ninguém soubesse de antemão quais seriam exatamente as implicações. Isso criaria uma

situação análoga ao “véu da ignorância” de Rawls.[350] Tal configuração poderia facilitar um consenso, ajudaria a evitar conflitos e promoveria um resultado mais igualitário.

Outra contraposição a alguns tipos de oráculos e gênios é que existem riscos no desenvolvimento de uma superinteligência cujo objetivo final não corresponda exatamente àquilo que nós procuramos obter como resultado final. Por exemplo, se usarmos uma motivação de domesticidade para fazer com que a superinteligência queira minimizar alguns de seus próprios impactos no mundo, poderíamos estar criando um sistema com um ranking de preferências em relação a possíveis resultados que seja diferente daquele de seu criador. O mesmo acontecerá se criarmos uma IA que dê um valor peculiarmente alto à sua capacidade de responder perguntas corretamente ou de obedecer fielmente a algum comando. Agora, se isso for tratado com cuidado, não deveriam haver maiores problemas: existiria concordância o suficiente entre os dois rankings — ao menos enquanto estiverem relacionados a mundos possíveis que tenham uma chance razoável de se tornar reais — para que os resultados que sejam bons de acordo com os padrões da IA também o sejam de acordo com os padrões do diretor. Mas talvez pudéssemos argumentar em favor do princípio de que seria insensato introduzir até mesmo uma quantidade limitada de desarmonia entre os objetivos da IA e os nossos. (É lógico que a mesma preocupação se aplicaria em relação ao fornecimento de objetivos a um soberano que não estejam em completa harmonia com os nossos.)

IAs ferramentas

Uma sugestão comum é aquela em favor da construção de superinteligências que sejam mais semelhantes a ferramentas do que a agentes.[351] Essa ideia parece surgir a partir da observação de que os softwares comuns, que são usados em inúmeras aplicações, não despertam qualquer tipo de preocupação remotamente análoga aos desafios discutidos neste livro. Não seria possível criar uma “IA ferramenta” que funcionasse como um desses softwares — semelhante, por exemplo, a um sistema de controle de voo ou a um assistente virtual —, porém mais flexível e capaz? Por que construir uma superinteligência que tenha vontade própria? Dentro dessa linha de raciocínio, o paradigma do agente é fundamentalmente enganoso. Em vez de criar uma IA que tenha crenças e desejos e que aja como uma pessoa artificial, nós deveríamos visar a criação de um software regular, que simplesmente faz o que é programado para fazer.

A ideia de criar software que “simplesmente faz o que é programado para fazer” não é, entretanto, tão óbvia quando o produto for uma inteligência geral poderosa. Existe, é claro, um senso comum segundo o qual todo software simplesmente faz o que é programado para fazer: seu comportamento é matematicamente especificado pelo código. Mas isso é igualmente verdade para todas as castas de inteligência de máquina, quer seja uma “IA ferramenta” ou não. Se, em vez disso, “simplesmente fazer o que é programado para fazer” significa que o software se comporta conforme as *pretensões* de seus programadores, então esse é um padrão que softwares comuns constantemente falham em alcançar.

Em virtude da capacidade limitada dos softwares contemporâneos (em comparação àquela de uma superinteligência de máquina), as consequências de tais falhas são gerenciáveis, variando de insignificantes até muito custosas, mas de nenhuma forma oferecendo um risco existencial.[352] Entretanto, se a insuficiência de capacidade, e não a suficiência de confiabilidade, faz com que os softwares comuns sejam existencialmente seguros, não é claro como esses softwares poderiam servir de modelo para uma superinteligência segura. É possível imaginar que a expansão do alcance das tarefas realizadas por um software comum poderia eliminar a necessidade de uma inteligência artificial geral. Mas o alcance e a diversidade de tarefas que uma inteligência geral poderia executar de forma lucrativa em uma economia moderna são enormes. Seria impraticável a criação de softwares de propósito específico para lidar com cada uma dessas tarefas. Mas mesmo que isso pudesse ser feito, a execução de tal projeto demandaria um *longo* tempo. Antes que pudesse ser completado, a natureza de algumas dessas tarefas já teria se modificado e novas tarefas já teriam se tornado relevantes. Existiria uma grande vantagem em possuir softwares que pudessem aprender novas tarefas por conta própria e, mais ainda, que pudessem descobrir novas tarefas que precisassem ser executadas. Mas, para isso, o software teria que ser capaz de aprender, raciocinar e planejar, e deveria fazê-lo de maneira robusta e poderosa em várias áreas do conhecimento. Em outras palavras, ele teria que possuir uma inteligência geral.

A tarefa de desenvolvimento de software em si é especialmente relevante para os nossos propósitos. Existiriam enormes vantagens práticas em se automatizar esse processo. Contudo, a capacidade para um autoaperfeiçoamento rápido é justamente a principal propriedade que possibilita a uma IA embrionária iniciar uma explosão de inteligência.

Se inteligência geral não é dispensável, existe alguma outra maneira de construir uma IA ferramenta de modo que sejam preservadas as qualidades passivas tranquilizadoras de uma ferramenta monótona? É possível obter uma inteligência geral que não seja um agente? Intuitivamente, não é apenas a capacidade limitada dos softwares comuns o que os torna seguros: é também sua falta de ambição. Não há nenhuma sub-rotina no Excel que desejaria secretamente dominar o mundo caso fosse inteligente o suficiente para fazê-lo. Uma aplicação de planilhas não “quer” nada; ela apenas executa cegamente as instruções contidas no programa. O que (alguém poderia se perguntar) se coloca no caminho da criação de uma aplicação do mesmo tipo com inteligência de caráter mais geral? Por exemplo, um oráculo ao qual fosse dada a descrição de um objetivo responderia com um plano de execução da mesma maneira que o Excel responde a uma coluna de números calculando sua soma — sem expressar quaisquer “preferências” a respeito do resultado ou em relação à maneira como os humanos irão utilizá-lo?

A maneira clássica de se escrever um software requer que o programador compreenda de forma suficientemente detalhada a tarefa que deverá ser executada para que ele seja capaz de formular um processo de solução explícito que consista em uma sequência de passos matematicamente bem-definidos e que possam ser escritos em linguagem de programação.[353] (Na prática, engenheiros de software

utilizam bibliotecas de códigos repletas de comportamentos úteis, que podem ser invocados sem a necessidade de entendimento de como esses comportamentos são implementados, mas esses códigos foram, por sua vez, criados por programadores que tinham uma compreensão detalhada a respeito do que estavam fazendo.) Essa abordagem funciona para a resolução de tarefas bem conhecidas e é utilizada na maioria dos softwares em uso hoje em dia. Entretanto, ela fica devendo quando ninguém sabe ao certo como solucionar todas as tarefas que precisarão ser executadas. É nesse ponto em que as técnicas do campo da inteligência artificial se tornam relevantes. Em aplicações restritas, a aprendizagem de máquina pode ser usada meramente para fazer um ajuste fino de alguns parâmetros em um programa desenvolvido majoritariamente por humanos. Um filtro de *spam*, por exemplo, pode ser treinado com base em um corpus de e-mails pré-classificados por meio da aplicação de um processo que modifica os pesos que o algoritmo de classificação atribui às várias características de diagnóstico. Em uma aplicação mais ambiciosa, o classificador poderia ser construído de forma que pudesse descobrir novas características por conta própria e testar sua validade em um ambiente que se modifica constantemente. Um filtro de *spam* ainda mais sofisticado poderia ser dotado com a habilidade de raciocinar a respeito das principais escolhas do usuário ou sobre o conteúdo das mensagens que ele está classificando. Em nenhum desses casos o programador precisa saber qual é a melhor maneira de distinguir o que é *spam* do que não é, mas apenas ser capaz de montar um algoritmo que consiga melhorar seu próprio desempenho através do aprendizado, da descoberta ou do raciocínio.

Com avanços em inteligência artificial, seria possível ao programador se livrar de uma parcela do trabalho cognitivo necessário para descobrir como realizar determinada tarefa. Em um caso extremo, o programador poderia simplesmente especificar um critério formal para o que contaria como sucesso e deixar a cargo da IA a busca por uma solução. A IA guiaria essa busca por meio de um conjunto poderoso de heurísticas e outros métodos para descobrir estruturas no espaço de possíveis soluções. Ela continuaria procurando até que encontrasse uma solução que se encaixasse no critério de sucesso. A IA, então, implementaria a solução ou (no caso de um oráculo) poderia informá-la ao usuário.

Formas rudimentares dessa abordagem são amplamente utilizadas hoje em dia. Entretanto, softwares que utilizam IA e técnicas de aprendizado de máquina, embora possuam alguma habilidade para encontrar soluções que os programadores não tenham antecipado, funcionam como ferramentas para todos os propósitos práticos, sem apresentar nenhum risco existencial. Nós entraríamos em uma zona de perigo apenas se esses métodos usados na busca de soluções se tornassem extremamente amplos e poderosos: ou seja, quando começarem a se equivaler a uma inteligência geral — e principalmente quando se equipararem a uma superinteligência.

Existem (pelo menos) dois lugares onde poderiam, então, surgir problemas. Primeiro, o processo de busca superinteligente pode encontrar uma solução que não é apenas inesperada como também totalmente não intencional. Isso poderia levar a uma falha de um dos tipos discutidos previamente (“instanciação perversa”,

“profusão de infraestrutura” ou “crimes da mente”). A maneira como isso poderia acontecer é mais óbvia no caso de um soberano ou de um gênio que implementem diretamente as soluções encontradas por eles. Se a criação de *emoticons* moleculares ou a transformação do planeta em clipes de papel for a primeira ideia encontrada por uma superinteligência que responde ao critério de sucesso, então teremos *emoticons* e clipes de papel.[354] Mas até mesmo um oráculo que — caso tudo desse certo — meramente *relatasse* a solução encontrada poderia também se tornar um caso de instanciação perversa. O usuário solicita ao oráculo um plano para alcançar um certo resultado ou uma tecnologia que sirva a um certo propósito, o que poderá resultar em uma instanciação perversa quando o usuário segue o plano ou constrói a tecnologia, da mesma maneira que ocorreria se a IA tivesse implementado a solução por ela mesma.[355]

Também poderiam surgir problemas ao longo do curso de operação do software. Se os métodos que o software utiliza na busca por uma solução forem suficientemente sofisticados, eles talvez incluam provisões para gerenciar o próprio processo de busca de maneira inteligente. Nesse caso, a máquina que estiver rodando o software poderá começar a se parecer menos com uma simples ferramenta e cada vez mais com um agente. Dessa forma, o software poderia começar desenvolvendo um plano com o objetivo de proceder em sua busca por uma solução. Esse plano poderia especificar quais áreas deveriam ser exploradas primeiro e por meio de quais métodos, quais dados deveriam ser coletados e como fazer o melhor uso possível dos recursos computacionais disponíveis. Na busca por um plano que satisfaça os critérios internos do software (como obter uma probabilidade suficientemente alta de encontrar uma solução que corresponda aos critérios especificados pelo usuário dentro do tempo terminado), o software pode se deparar com ideias pouco ortodoxas. Por exemplo, ele pode elaborar um plano que comece com a aquisição de recursos computacionais adicionais e a eliminação de potenciais fontes de interrupção (como seres humanos). Tais planos “criativos” surgiriam quando as habilidades cognitivas do software atingissem um nível suficientemente alto. Quando o software colocasse tal plano em prática, o resultado poderia ser uma catástrofe existencial.

Conforme ilustrado pelos exemplos do quadro 9, processos de busca abertos às vezes evidenciam soluções não antropocêntricas, estranhas e inesperadas, mesmo em suas atuais formas limitadas. Os processos de busca disponíveis atualmente não apresentam perigo porque são muito fracos para descobrir o tipo de plano que poderia capacitar um programa a dominar o mundo. Tal plano incluiria etapas extremamente complicadas, como a invenção de uma nova tecnologia armamentista que estivesse muitas gerações à frente da tecnologia mais avançada ou a execução de uma campanha de propaganda muito mais efetiva do que as desenvolvidas pelos melhores publicitários humanos. Para conseguir pelo menos *conceber* essas ideias, quem dirá desenvolvê-las de modo que funcionassem efetivamente, uma máquina provavelmente precisaria ter a capacidade de representar o mundo de uma maneira que fosse ao menos tão rica e realista quanto a representação de mundo que um humano adulto normal possui (embora a falta de conhecimento em algumas áreas pudesse ser compensada por habilidades extras

em outros campos). Isso está muito além do alcance da IA contemporânea. E por conta da explosão combinatória, que geralmente frustra as tentativas de solucionar os problemas de planejamento complicados na base da força bruta (como vimos no capítulo 1), as deficiências dos algoritmos conhecidos não podem ser realisticamente superadas simplesmente com o uso de mais poder computacional.[356] Entretanto, uma vez que a busca e os processos de planejamento se tornem suficientemente poderosos, também se tornarão potencialmente perigosos.

Em vez de tentar fazer com que o comportamento de um agente dotado de propósito surja casualmente de forma espontânea a partir da implementação de poderosos processos de busca (incluindo processos que busquem planos de ação internos e processos que busquem diretamente soluções que satisfaçam algum critério definido pelo usuário), talvez seja melhor criar agentes de maneira intencional. Dotar uma superinteligência com uma estrutura explicitamente semelhante à de um agente pode ser uma maneira de aumentar sua previsibilidade e transparência. Um sistema bem projetado, construído de modo que exista uma separação clara entre seus valores e suas crenças, nos permitiria prever alguns dos resultados que ele tenderia a produzir. Mesmo que não conseguíssemos antever exatamente quais crenças o sistema teria ou em que situações ele se encontraria, saberíamos exatamente para onde ir a fim de inspecionar seus valores finais e, portanto, os critérios que ele usaria em seus planos de ação futuros ou na avaliação de qualquer plano em potencial.

Comparação

É útil resumir as características das diferentes castas de sistema que discutimos aqui (tabela 11).

Mais pesquisas serão necessárias para determinar qual tipo de sistema seria o mais seguro. A resposta pode depender das condições sob as quais a IA seria utilizada. A casta dos oráculos é obviamente atrativa do ponto de vista de segurança, uma vez que permitiria a aplicação tanto de métodos de controle de capacidade quanto de métodos de seleção de motivação. Por esse motivo, ela pode parecer mais vantajosa que a casta dos soberanos, que permitiria apenas a aplicação de métodos de seleção de motivação (exceto em cenários onde se acredite que o mundo possa conter outras superinteligências poderosas, nos quais integração social e captura antrópica poderiam ser aplicáveis). Entretanto, um oráculo seria capaz de colocar um enorme poder nas mãos do seu operador, o qual poderia ser corrompido ou aplicá-lo de modo irresponsável, enquanto um soberano ofereceria alguma proteção contra esses perigos. Sendo assim, o ranking de segurança não é algo simples de ser determinado.

Um gênio pode ser visto como um meio-termo entre um oráculo e um soberano — mas não necessariamente um bom meio-termo. Ele compartilharia, em muitos aspectos, das desvantagens de ambos. A aparente segurança de uma IA ferramenta pode ser ilusória. Para que as ferramentas sejam versáteis o suficiente para substituir agentes superinteligentes, elas poderiam demandar a aplicação de

processos internos de busca e planejamento extremamente poderosos. Comportamentos semelhantes aos de um agente podem surgir a partir desses processos como uma consequência não planejada. Nesse caso, seria melhor desenvolver o sistema de maneira que ele fosse deliberadamente um agente, e assim os programadores poderiam visualizar mais facilmente quais critérios acabariam por determinar os resultados obtidos por tal sistema.

QUADRO 9

Soluções estranhas em uma busca cega

Mesmo processos de busca simples e evolutivos às vezes produzem resultados bastante inesperados, na forma de soluções que satisfazem um critério formal definido pelo usuário de um modo muito diferente do que esse usuário pretendia ou esperava.

O campo de estudo dos hardwares evolutivos oferece muitos exemplos desse fenômeno. Nesse campo, um algoritmo evolutivo faz uma busca no espaço de possíveis projetos de hardware, realiza testes de aptidão em cada um desses projetos e os implementa fisicamente em uma placa que possa ser rapidamente configurável. Os designs "evoluídos" demonstram, em geral, uma extraordinária economia. Por exemplo, uma busca descobriu um circuito de discriminação de frequências que funciona sem um relógio — um componente geralmente considerado necessário para essa função. Os pesquisadores estimaram que o circuito evoluído era entre uma e duas ordens de magnitude menor do que um engenheiro humano teria projetado para a tarefa. O circuito tirava partido das propriedades físicas de seus componentes de maneiras não ortodoxas; alguns componentes ativos e necessários não foram sequer conectados aos pinos de entrada e saída! Em vez disso, esses componentes participavam por uma via que seria considerada um efeito colateral inconveniente, como acoplamento eletromagnético ou carregamento de energia.

Outro processo de busca, encarregado de criar um oscilador, foi privado de um componente aparentemente ainda mais indispensável: o capacitor. Quando o algoritmo apresentou uma solução mais bem-sucedida, os pesquisadores a examinaram e concluíram primeiramente que ela "não deveria funcionar". Após uma análise mais cuidadosa, eles descobriram que o algoritmo tinha conseguido, num estilo MacGyver, reconfigurar sua placa-mãe, que não tinha sensores, para que funcionasse como um receptor improvisado de sinais de rádio, usando as trilhas da placa de circuito impresso como uma antena para captar os sinais gerados por computadores pessoais que, por acaso, estavam localizados perto do laboratório. O circuito amplificou esse sinal para gerar a saída oscilatória desejada.[357]

Em outros experimentos, algoritmos evolutivos projetaram circuitos capazes de perceber se a placa-mãe estava sendo monitorada com um osciloscópio ou se um ferro de solda tinha sido conectado à fonte de energia comum do laboratório. Esses exemplos ilustram como um processo de busca aberta pode adaptar materiais acessíveis para criar capacidades sensoriais completamente inesperadas, de maneiras que o pensamento humano convencional não estaria apropriadamente preparado para tirar proveito ou sequer levar em consideração.

A tendência que as buscas evolutivas apresentam de "trapacear" ou encontrar soluções contraintuitivas para um determinado problema também pode ser percebida na natureza, embora talvez seja algo menos óbvio para nós, pois estamos até certo ponto familiarizados com o funcionamento e a aparência da biologia, de modo que nos vemos propensos a considerar os resultados de processos evolutivos naturais como algo normal — mesmo que não esperássemos por eles *ex ante*. Mas é possível criar experimentos em seleção artificial, nos quais podemos ver o processo evolutivo em ação, fora de seu contexto familiar. Em tais experimentos, pesquisadores podem criar condições que raramente são obtidas na natureza para observar seus resultados.

Por exemplo, antes da década de 1960, era aparentemente muito comum biólogos afirmarem que populações de predadores restringiam sua própria procriação para evitar cair em uma armadilha malthusiana.[358] Embora a seleção individual trabalhe contra essa restrição, por algum tempo foi considerado que a seleção de grupo se sobreporia a incentivos individuais para se tirar proveito de oportunidades de reprodução, favorecendo traços que poderiam beneficiar o grupo ou a população em geral. Mais tarde, análises teóricas e estudos baseados em simulações mostraram que, embora a seleção de grupo seja, em princípio, possível, ela pode se sobrepor à forte seleção individual apenas sob condições muito restritas que dificilmente ocorrem na natureza.[359] Mas tais condições podem ser criadas em um laboratório. Quando besouros-castanhos (*Tribolium castaneum*) foram cruzados com o intuito de reduzir o tamanho da sua população por meio da aplicação de uma seleção de grupo forte, a evolução de fato levou a populações menores.[360] Entretanto, os meios pelos quais isso foi conseguido não incluíram apenas adaptações "benignas", como redução da fecundidade e aumento no período de desenvolvimento, algo que um humano que ingenuamente

antropomorfizasse o processo de busca evolutiva poderia esperar, mas também um aumento do canibalismo.[361]

TABELA 11

Características das diferentes castas de sistema

Oráculo	**Um sistema que responde perguntas** *Variações*: Oráculos de domínio limitado (por exemplo, matemática); oráculos com respostas limitadas (por exemplo, respostas apenas do tipo sim/não/indeterminado ou probabilidades); oráculos que se recusam a responder perguntas caso eles prevejam que as consequências de fornecer respostas se enquadrariam em "critérios de desastre" predefinidos; diversos oráculos que revisem as respostas um dos outros	•Métodos de confinamento plenamente aplicáveis •Domesticidade plenamente aplicável •Redução da necessidade de que a IA entenda as intenções e interesses humanos (em comparação com um gênio ou um soberano) •Uso de perguntas com respostas do tipo sim/não pode evitar a necessidade de uma métrica que avalie a "utilidade" ou o "valor informativo" das respostas •Fonte de grande poder (pode fornecer ao operador uma vantagem estratégica decisiva) •Proteção limitada contra o uso irresponsável por parte do operador •Oráculos pouco confiáveis poderiam ser usados para prover respostas difíceis de serem encontradas, mas de fácil verificação • Um sistema fraco de verificação de respostas poderia ser criado através do uso de múltiplos oráculos
Gênio	**Um sistema de execução de comandos** *Variações*: Gênios usando diferentes "distâncias de extrapolação" ou que sigam em diferentes graus o espírito dos comandos em vez de segui-los ao pé da letra; gênios de domínio restrito; gênios com a possibilidade de pré-visualização; gênios que se recusam a obedecer a comandos caso prevejam que as consequências de lhes obedecê-los se enquadrariam em "critérios de desastre" predefinidos	•Métodos de confinamento parcialmente aplicáveis (para gênios espacialmente limitados) •Domesticidade parcialmente aplicável •Um gênio pode fornecer uma previsão de aspectos relevantes dos resultados esperados •Um gênio pode implementar mudanças de modo gradual, permitindo oportunidades de revisão em cada fase •Fonte de grande poder (pode fornecer ao operador uma vantagem estratégica decisiva) •Proteção limitada contra o uso irresponsável por parte do operador

		• Maior necessidade de que a IA entenda interesses e intenções humanas (em comparação aos oráculos)
Soberano	**Um sistema projetado para a execução autônoma de operações indeterminadas** *Variações*: Muitos sistemas de motivação possíveis; possibilidade de pré-visualização e "ratificação pelo patrocinador" (conceito que será discutido no capítulo 13)	•Métodos de confinamento não são aplicáveis •A maioria dos outros métodos de controle de capacidades também não são aplicáveis (exceto, talvez, integração social ou captura antrópica) •Domesticidade não é aplicável na maioria dos casos •Maior necessidade de que a IA entenda os verdadeiros interesses e intenções humanas •É necessário que seja criado de forma correta na primeira tentativa (embora, possivelmente em um grau menor, isso também seja verdadeiro para as outras castas) •Potencial fonte de grande poder para o patrocinador, incluindo uma vantagem estratégica decisiva •Uma vez ativado, se torna invulnerável ao controle por parte do operador e pode ser desenvolvido com algum tipo de proteção contra uso irresponsável • Pode ser usado para implementar resultados do tipo "véu da ignorância" (cf. capítulo 13)
Ferramenta	**Um sistema que não tenha sido desenvolvido para exibir comportamentos direcionados à execução de objetivos**	•Métodos de confinamento podem ser aplicáveis, dependendo da implementação •Processos poderosos de busca estariam provavelmente envolvidos no desenvolvimento e operação de uma superinteligência de máquina •Processos poderosos de busca para encontrar soluções que se encaixem em algum critério formal podem produzir soluções que se adéquem aos critérios de uma maneira não intencional e perigosa • Processos de busca poderosos podem envolver processos secundários internos de busca e planejamento que possivelmente encontrariam

meios perigosos de executar o processo principal de busca

11

CAPÍTULO 11

CENÁRIOS MULTIPOLARES

Vimos (particularmente no capítulo 8) o quanto um desfecho unipolar, no qual uma única superinteligência viesse a obter uma vantagem estratégica decisiva e a utilizasse para criar um *singleton*, poderia ser ameaçador. Neste capítulo, examinaremos o que poderia acontecer em um desfecho multipolar, ou seja, uma sociedade pós-transição com múltiplas agências superinteligentes competindo entre si. Nosso interesse nessa classe de cenários é duplo. Primeiro, conforme mencionado no capítulo 9, pode-se imaginar que a integração social ofereça uma solução para o problema de controle. Já citamos algumas das limitações de tal abordagem, e este capítulo procurará apresentar uma visão mais completa. Segundo, ainda que ninguém se dedicasse a criar condições para um desfecho multipolar como um meio de lidar com o problema de controle, esse desfecho poderá ocorrer de qualquer maneira. E com o que, então, esse desfecho se pareceria? A sociedade competitiva que resultaria dele não é necessariamente atrativa e tampouco duradoura.

Em cenários que envolvem um *singleton*, o que aconteceria após a transição dependeria quase que inteiramente dos valores desse *singleton*. O desfecho poderia ser, dessa forma, muito bom ou muito ruim, dependendo desses valores. E esses valores dependeriam, por sua vez, da eventual resolução do problema do controle — e em que medida isso ocorreu — e dos objetivos do projeto que criou o *singleton*.

Os interessados no desfecho dos cenários que envolvem um *singleton* teriam apenas três fontes de informação: informação acerca de coisas que não podem ser afetadas pelas ações do *singleton* (tais como as leis da física); informação acerca dos valores instrumentais convergentes; e informação que nos permite prever ou especular a respeito dos valores finais que o *singleton* terá.

Em cenários multipolares, um conjunto adicional de restrições entra em jogo, e tais restrições se relacionam com a forma como os agentes interagem entre si. As dinâmicas sociais que emergem dessas interações podem ser estudadas por meio de técnicas oriundas da teoria dos jogos, da economia e da teoria da evolução. Elementos de ciência política e sociologia também são relevantes na medida em que podem ser extraídos e abstraídos de características mais contingentes da experiência humana. Embora seja pouco realista esperarmos que essas restrições nos deem uma imagem precisa do mundo pós-transição, elas podem nos ajudar a identificar algumas possibilidades relevantes e a contestar algumas hipóteses injustificadas.

Iniciaremos com a análise de um cenário econômico caracterizado por um nível baixo de regulação, forte proteção ao direito de propriedade e uma introdução relativamente rápida de mentes digitais de baixo custo.[362] Esse tipo de modelo é associado especialmente ao economista norte-americano Robin Hanson, responsável por estudos pioneiros sobre o assunto. Mais adiante neste capítulo, nos concentraremos em algumas considerações evolucionárias e examinaremos as possibilidades de um mundo pós-transição inicialmente multipolar que fosse posteriormente unificado em um *singleton*.

Sobre cavalos e homens

A inteligência geral de máquina poderia servir como um substituto para a inteligência humana. As mentes digitais poderiam não somente executar o trabalho intelectual realizado atualmente pelos humanos, mas, uma vez equipadas com bons atuadores ou corpos robóticos, as máquinas poderiam também substituir o trabalho braçal realizado pelos humanos. Suponha que trabalhadores-máquinas — que poderiam ser rapidamente reproduzidos — se tornassem mais baratos e mais capazes que os trabalhadores humanos em praticamente todos os tipos de trabalho. O que aconteceria a partir daí?

Salários e desemprego

Com a possibilidade de reproduzir o trabalho a um baixo custo, os salários no mercado despencariam. Os únicos setores nos quais os humanos permaneceriam competitivos seriam aqueles onde os clientes tivessem preferência por serviços realizados por humanos. Atualmente, bens manufaturados ou produzidos por povos indígenas muitas vezes possuem preços mais elevados. No futuro, os consumidores poderão, igualmente, preferir produtos que tenham sido feitos por humanos, assim como atletas, artistas, amantes e líderes humanos, em vez de equivalentes artificiais, funcionalmente indistinguíveis ou superiores. Não está claro, no entanto, quão comuns seriam essas preferências. Caso os produtos e serviços oferecidos pelas máquinas fossem suficientemente superiores, eles poderiam ser mais valorizados.

Um parâmetro que poderá ser relevante para a escolha do consumidor é a vida interior do trabalhador que oferece um serviço ou um produto. O público de um concerto, por exemplo, poderá querer saber que o intérprete está experimentando a música e o palco de forma consciente. Na ausência dessa “alma”, o músico poderia ser meramente visto como uma caixa de música de alta potência, embora fosse capaz de criar uma imagem tridimensional de um intérprete apto a interagir naturalmente com a plateia. Máquinas poderiam então ser projetadas para instanciar os mesmos tipos de estados mentais que estariam presentes num humano que executasse a mesma tarefa. Entretanto, mesmo com a replicação perfeita de experiências subjetivas, algumas pessoas poderiam simplesmente preferir o trabalho orgânico. Tais preferências poderiam ter razões ideológicas ou religiosas. Assim como muitos muçulmanos e judeus evitam comidas preparadas de formas que eles classificam como *haram* ou *treif*, é provável que no futuro venham a

existir grupos que evitem produtos cuja manufatura envolva o uso não autorizado de inteligência de máquina.

E o que poderia tornar isso possível? Na medida em que o trabalho barato das máquinas consiga substituir o trabalho humano, postos de trabalho antes ocupados por humanos poderiam desaparecer. É claro que a apreensão causada pela automatização e perda de empregos não é nova. A preocupação acerca do desemprego causado pela tecnologia tem vindo à tona periodicamente pelo menos desde a Revolução Industrial; e algumas profissões, de fato, seguiram o caminho dos tecelões e artesãos ingleses que, unidos sob a bandeira do folclórico "general Ludd", lutaram contra a introdução dos teares mecânicos no início do século XIX. No entanto, embora a tecnologia e os maquinários venham substituindo uma série de tipos específicos de trabalho humano, a tecnologia física tem sido, de uma forma geral, complementar ao trabalho. Os salários médios de humanos ao redor do mundo têm apresentado uma tendência de crescimento ao longo do tempo, em grande parte por causa dessas complementaridades. Ainda assim, o que começa como um complemento ao trabalho pode, num estágio mais avançado, substituí-lo. Os cavalos foram inicialmente complementados com carroças e arados, o que aumentou, em muito, a produtividade desses animais. Mais tarde, os cavalos foram substituídos por automóveis e tratores. Essas inovações reduziram a demanda por trabalho equino e ocasionaram um colapso dessa população. A espécie humana poderia estar sujeita a um destino semelhante?

Podemos aprofundar esse paralelo com a história dos cavalos se nos perguntarmos por que ainda há cavalos por aí. Uma das razões é o fato de ainda existirem alguns nichos nos quais os cavalos possuem vantagens funcionais, como no trabalho policial. Mas a principal razão é o fato de que os humanos têm preferências peculiares pelos serviços que os cavalos podem realizar, incluindo serviços de recreação, como a equitação e a corrida de cavalos. Essas preferências podem ser comparadas àquelas que no futuro alguns humanos poderiam ter, como a preferência por certos produtos e serviços produzidos artesanalmente por humanos. Embora sugestiva, essa analogia é inexata, já que não há ainda um substituto funcional completo para os cavalos. Caso existissem aparelhos mecânicos baratos que utilizassem o feno como combustível e tivessem o mesmo formato, cheiro e comportamento que os cavalos biológicos — e talvez até as mesmas experiências conscientes —, então a demanda por cavalos biológicos provavelmente diminuiria ainda mais.

Com uma redução suficiente na demanda pela mão de obra humana, os salários cairiam abaixo do nível de subsistência humana. Os potenciais inconvenientes para os trabalhadores humanos seriam, então, extremos: não apenas cortes de salário, rebaixamentos ou necessidade de requalificação, mas também fome e morte. Quando os cavalos se tornaram obsoletos como força de trabalho para transporte, muitos foram vendidos para frigoríficos e transformados em ração para cachorros, farinha de osso, couro e cola. Esses animais não tiveram empregos alternativos que justificassem sua manutenção. Nos Estados Unidos havia em torno de 26 milhões de cavalos em 1915. No início da década de 1950 existiam apenas 2 milhões.[363]

Capital e assistência social

Uma diferença entre os cavalos e os humanos é que os humanos possuem capital. É um fato empírico que a participação total do capital tem se mantido estável ao redor dos 30% há bastante tempo (apesar de significativas flutuações de curto prazo).[364] Isso significa que 30% da renda global total é recebida na forma de rendimentos pelos donos do capital, enquanto os 70% restantes são recebidos pelos trabalhadores em forma de salários. Se classificarmos uma IA como capital, então, com a invenção de uma inteligência de máquina que possa substituir integralmente a mão de obra humana, os salários se equivaleriam ao reduzido valor do custo de tais máquinas e — considerando que as máquinas seriam muito eficientes — seriam bem baixos, muito menores do que a renda necessária para a subsistência humana. A proporção da renda recebida por meio do trabalho diminuiria, chegando praticamente a zero. Mas isso sugere que a proporção do capital ficaria próxima de 100% do total do PIB mundial. Uma vez que o produto interno bruto mundial dispararia após uma explosão de inteligência (em virtude do número enorme de máquinas capazes de substituir trabalhadores, mas também por causa da conquista, pela superinteligência, de avanços tecnológicos e, posteriormente, por conta da aquisição de vastas quantidades de terras através da colonização do espaço), o resultado seria um enorme aumento na renda total do capital. Se os humanos permanecessem como donos desse capital, o total da renda recebida pela população humana aumentaria astronomicamente, embora, nesse cenário, os humanos não mais receberiam qualquer tipo de salário.

A espécie humana como um todo poderia então se tornar rica de uma forma inimaginável. Como essa renda seria distribuída? Numa primeira aproximação, o lucro do capital seria proporcional ao montante do capital possuído. Dado o efeito de amplificação astronômico, mesmo uma pequena porção de patrimônio no período pré-transição se transformaria numa imensa fortuna no período pós-transição. Entretanto, no mundo contemporâneo, muitas pessoas não possuem patrimônio algum. Isso inclui não apenas quem vive na pobreza, mas também algumas pessoas que recebem uma boa renda ou que possuem grande capital humano, embora tenham um patrimônio líquido negativo. Por exemplo, em países prósperos como a Dinamarca e a Suíça, 30% da população declara patrimônio negativo — geralmente pessoas jovens, de classe média, com poucos ativos tangíveis e com dívidas de cartão de crédito ou de financiamento estudantil.[365] Mesmo que as poupanças pudessem render juros extremamente altos, seria necessário o investimento de algum capital inicial para dar início ao rendimento.[366]

Apesar disso, até mesmo indivíduos que não possuíssem nenhum capital privado no início da transição poderiam se tornar extremamente ricos. Aqueles que participassem de um plano de pensão, público ou privado, por exemplo, estariam numa boa posição, contanto que esse plano fosse, ao menos parcialmente, financiado.[367] Os mais pobres também poderiam enriquecer por meio da filantropia praticada por aqueles que vissem seus patrimônios líquidos crescerem de forma excepcional: em virtude do tamanho astronômico da prosperidade, mesmo doações que correspondessem a uma fração muito pequena seriam equivalentes a quantias muito grandes em termos absolutos.

Também é possível que as pessoas ainda pudessem enriquecer através do trabalho, mesmo no período pós-transição, quando as máquinas fossem funcionalmente superiores aos humanos em todas as áreas (e também mais baratas que o trabalho humano em nível de subsistência). Como já dissemos, isso poderia acontecer se houvesse nichos nos quais o trabalho realizado por humanos fosse preferido por razões estéticas, ideológicas, éticas, religiosas ou por outros motivos não pragmáticos. Em um cenário no qual o número de humanos detentores de capital aumentasse drasticamente, a demanda por trabalho humano poderia crescer de forma correspondente. Os novos trilionários e quatrilionários poderiam se dar ao luxo de pagar grandes somas para ter alguns de seus bens e serviços oferecidos pela força de trabalho orgânica, no estilo *fair trade*. A história dos cavalos nos oferece novamente um paralelo. Depois que a população de cavalos nos Estados Unidos caiu para 2 milhões no início dos anos 1950, houve uma recuperação robusta: um censo recente apontou a existência de menos de 10 milhões de cavalos.[368] Esse aumento não se deve à nenhuma nova necessidade funcional de cavalos na agricultura ou nos transportes, mas ao crescimento econômico que possibilitou que mais norte-americanos pudessem pagar pela recreação equestre.

Outra diferença relevante existente entre cavalos e humanos, além da capacidade humana de deter a propriedade do capital, é a de que os humanos são capazes de se mobilizar politicamente. Um governo exercido por humanos poderia usar o poder que o Estado possui de cobrar impostos para redistribuir lucros privados, ou aumentar a receita através da venda de patrimônios estatais valiosos, como terras públicas, e usar essa receita para pagar pensões aos seus constituintes. Novamente, em virtude do crescimento econômico explosivo, durante e imediatamente após a transição, haveria muitíssimo mais riquezas disponíveis, tornando relativamente fácil a tarefa de colocar comida no prato de todos os cidadãos desempregados. Isso seria factível até para um único país que quisesse prover todos os humanos ao redor do mundo com um generoso salário de subsistência a um custo proporcional que não seria superior aos gastos com ajuda humanitária realizada por muitos países atualmente.[369]

O princípio malthusiano em uma perspectiva histórica

Até agora consideramos a população humana como um fator constante. Essa pode ser uma premissa aceitável quando lidamos com prazos curtos, já que a biologia limita a taxa de reprodução humana. Para prazos mais longos, no entanto, essa premissa não é necessariamente razoável.

A população humana aumentou mil vezes nos últimos 9 mil anos.[370] Se a população humana não tivesse se visto atrelada aos limites impostos pela economia mundial durante toda a pré-história e a história, esse crescimento poderia ter ocorrido muito mais rapidamente. Uma condição aproximadamente malthusiana prevaleceu, na qual a maioria das pessoas recebia rendas que apenas permitiam sua sobrevivência e a criação, em média, de dois filhos até a fase adulta.[371] Havia atribulações locais e temporárias: pragas, variações climáticas e guerras que intermitentemente diminuíam a população e liberavam terra, permitindo que os

sobreviventes melhorassem sua ingestão nutricional — gerando mais crianças até que a população se refazia e o quadro malthusiano era reinstituído. Graças também à desigualdade social, uma pequena elite podia desfrutar de uma renda consistentemente acima do nível de subsistência (ao custo da diminuição do tamanho total da população que poderia ser sustentada). Temos aqui um pensamento triste e dissonante: que nesse quadro malthusiano, o cenário normal durante a maior parte de nossa existência neste planeta foi marcado por secas, pestes, massacres e desigualdade — os piores adversários do bem-estar humano de acordo com as avaliações mais comuns —, que podem também ter sido os maiores humanitários, pois apenas esses fatores permitiram que o nível médio de bem-estar aumentasse levemente acima do nível de subsistência.

Sobreposto às flutuações locais, a história apresenta um padrão macro de crescimento econômico inicialmente lento, mas que vem acelerando, alimentado pela acumulação de inovações tecnológicas. O crescimento da economia mundial trouxe consigo um aumento proporcional da população global. (Mais precisamente, uma população maior parece ter acelerado consideravelmente a taxa de crescimento econômico, talvez principalmente em virtude do aumento da inteligência coletiva da humanidade.)[372] No entanto, somente depois da Revolução Industrial o crescimento econômico se tornou tão rápido que o crescimento da população não conseguiu acompanhá-lo. Dessa forma, a renda média começou a aumentar, primeiro nos países industrializados da Europa ocidental e, em seguida, na maior parte do mundo. Atualmente, mesmo nos países mais pobres, a renda média excede substancialmente o nível mínimo de subsistência, o que se reflete no fato de que as populações desses países vêm crescendo.

Os países mais pobres apresentam atualmente as maiores taxas de crescimento populacional, pois eles ainda não concluíram a “transição demográfica” para um regime de baixa natalidade como aconteceu nas sociedades mais desenvolvidas. Os demógrafos projetam que a população mundial crescerá até atingir 9 bilhões em meados deste século e que poderá, a partir daí, estacionar ou diminuir conforme os países mais pobres se juntem às nações desenvolvidas nesse regime de baixa fertilidade.[373] Muitos países ricos já possuem taxas de fertilidade que estão abaixo do nível necessário para substituição; em alguns casos, muito abaixo.[374]

Ainda assim há razões, se tivermos uma visão de longo prazo e considerarmos um estado de tecnologia e prosperidade constantes, para esperarmos um retorno à condição histórica e ecologicamente normal de uma população mundial que se mantenha próxima aos limites que nosso nicho consegue suportar. Se isso parece contraintuitivo diante da relação negativa entre riqueza e fertilidade que estamos observando atualmente numa escala global, precisamos nos lembrar de que essa era moderna é apenas uma breve fração da história e uma grande aberração. O comportamento humano ainda não se adaptou às condições contemporâneas. Não apenas deixamos de tirar vantagem de formas óbvias para aumentar nossa aptidão de maneira mais abrangente (por exemplo, nos tornando doadores de óvulos ou espermas), mas sabotamos ativamente nossa fertilidade com o uso de métodos contraceptivos. Num ambiente de adaptabilidade evolucionária, um impulso sexual

saudável poderia, por si só, ser suficiente para fazer um indivíduo agir de modo a maximizar seu potencial reprodutivo; num ambiente moderno, no entanto, haveria uma enorme vantagem seletiva de se possuir um desejo mais direto de ser pai ou mãe biológica do maior número de crianças possível. Tal desejo vem sendo selecionado evolutivamente, em conjunto com outras características que aumentam nossa propensão à reprodução. No entanto, a adaptação cultural pode passar a perna na evolução biológica. Algumas comunidades, tais como a dos huteritas ou a dos adeptos do movimento evangélico Quiverfull, possuem culturas de natalidade que encorajam famílias numerosas e, dessa forma, estão se expandindo rapidamente.

Crescimento populacional e investimento

Se imaginarmos as atuais condições socioeconômicas congeladas magicamente na forma como conhecemos hoje, o futuro seria dominado por grupos étnicos e culturais que mantivessem altos níveis de fertilidade. Se a maioria das pessoas tivesse preferências que maximizassem sua aptidão no ambiente contemporâneo, a população poderia facilmente dobrar a cada geração. Na ausência de políticas de controle de natalidade — as quais teriam que se tornar constantemente mais rigorosas e efetivas para contrabalançar a evolução de preferências mais fortes de burlar essas políticas —, a população mundial, então, continuaria a crescer exponencialmente até que alguma restrição, como a escassez de terras ou o esgotamento de oportunidades fáceis para inovações significativas, tornasse impossível a manutenção do crescimento econômico nessa mesma velocidade: nesse ponto, a renda média começaria a diminuir até que alcançasse um nível em que a pobreza esmagadora impedisse a maioria das pessoas de criar mais do que duas crianças até a fase adulta. Desse modo, o princípio malthusiano se reafirmaria, como um temido senhor de escravos, dando fim à nossa fuga para uma terra de sonhos de abundância e nos levando de volta à pedreira, presos a correntes para retomar nossa desgastante luta pela subsistência.

Essa perspectiva de longo prazo poderia ser acelerada e se tornar mais iminente com a explosão de inteligência. Como um software pode ser copiado, uma população de emulados ou IAs poderia dobrar rapidamente — dentro de alguns minutos em vez de décadas ou séculos — e logo todos os hardwares disponíveis estariam esgotados.

A propriedade privada poderia oferecer proteção parcial contra o surgimento de um quadro malthusiano universal. Considere um modelo simples, onde clãs (comunidades fechadas ou Estados) iniciassem sua trajetória com quantidades variadas de propriedade e que adotassem de forma independente diferentes políticas de reprodução e investimento. Alguns clãs descontariam acentuadamente o valor de recursos no futuro e gastariam praticamente tudo o que possuíssem, então seus membros empobrecidos se juntariam ao proletariado global (ou morreriam, caso não fossem capazes de se manter através do seu trabalho). Outros clãs investiriam parte dos seus recursos, mas adotariam uma política de reprodução ilimitada: tais clãs se tornariam mais populosos até alcançar um quadro malthusiano interno, no qual seus membros seriam tão pobres que morreriam quase na mesma proporção de sua taxa de reprodução, ponto no qual o aumento da população do

clã diminuiria para se equiparar ao crescimento dos seus recursos. Já outros clãs poderiam restringir sua fertilidade a um nível inferior à taxa de crescimento do capital: tais clãs poderiam incrementar lentamente seus números enquanto a renda per capita aumentasse e seus membros fossem se tornando mais ricos.

Se a riqueza fosse redistribuída dos clãs mais ricos para os membros dos clãs que estivessem se reproduzindo mais rapidamente ou realizando um acentuado desconto do valor de recursos no futuro (cujas crianças, cópias ou descendentes, mesmo sem ter culpa nenhuma, seriam lançados no mundo com capital insuficiente para sobreviver e prosperar), então algo muito próximo a um quadro malthusiano universal ocorreria. Em um caso limite, todos os membros de todos os clãs receberiam renda suficiente para sobreviver e todos seriam iguais na pobreza.

Se a propriedade não fosse redistribuída, clãs prudentes poderiam reter uma certa quantidade de capital, e é possível que sua riqueza apresentasse um aumento em termos absolutos. Não é claro, entretanto, se os humanos seriam capazes de receber taxas de retorno tão altas pelo seu capital quanto as inteligências de máquina, pois pode ser que haja sinergias entre trabalho e capital, de modo que um único agente capaz de suprir a ambos (por exemplo, um empreendedor ou um investidor que seja ao mesmo tempo muito capacitado e muito rico) poderia obter uma taxa de retorno privada pelo seu capital que excederia a taxa de mercado obtida por agentes que possuíssem recursos financeiros, embora fossem destituídos de recursos cognitivos. Os humanos, sendo menos capacitados que as inteligências de máquina, poderiam, portanto, aumentar seu capital mais lentamente — a menos, é claro, que o problema do controle tenha sido resolvido por completo. Nesse caso, a taxa de retorno dos humanos se igualaria à taxa de retorno da máquina, já que um diretor humano poderia incumbir a um agente de máquina a administração de suas poupanças e seria capaz de fazê-lo sem custos ou conflitos de interesses. Caso contrário, a fração da economia controlada por máquinas se aproximaria assintoticamente dos 100%.

Um cenário no qual a fração da economia que é controlada por máquinas se aproxima assintoticamente dos 100% não é, necessariamente, uma ocasião em que a fatia humana da economia diminuiria. Se a economia crescesse numa taxa suficientemente alta, então mesmo uma fração relativamente decrescente dessa economia poderia, ainda assim, estar aumentando seu tamanho absoluto. Esta pode parecer uma notícia relativamente boa para a humanidade: em um cenário multipolar no qual os direitos de propriedade são protegidos — mesmo que falhemos completamente em resolver o problema do controle —, o volume total de riqueza dos humanos poderia aumentar. É claro que esse efeito não acabaria com o problema do crescimento da população humana e, por consequência, não solucionaria a questão da redução da renda per capita a níveis de subsistência. Também não resolveria o problema dos humanos que se arruinariam por descontar o valor dos recursos no futuro.

No longo prazo, a economia seria cada vez mais dominada pelos clãs que tivessem as maiores taxas de poupança — poupadores avarentos que seriam donos de metade da cidade, mas viveriam embaixo da ponte. Apenas na plenitude do tempo, quando não houvesse mais oportunidades de investimento, é que os

poupadores avarentos mais prósperos começariam a utilizar suas poupanças.[375] Entretanto, se a proteção aos direitos de propriedade não for perfeita — por exemplo, se as máquinas mais eficientes conectadas à rede conseguirem transferir as riquezas dos humanos para elas próprias, seja lá de que maneira —, então os capitalistas humanos podem precisar gastar o seu capital muito mais cedo, antes que seja dilapidado por tais transferências (ou pelos custos gerados pela necessidade de proteção contra essas transferências). Caso esses desenvolvimentos ocorram em uma escala de tempo digital em vez de em uma escala biológica, os humanos glaciais seriam completamente expropriados antes que pudessem perceber.[376]

A vida em uma economia algorítmica

A vida para os humanos biológicos num estágio malthusiano pós-transição não precisa se parecer com nenhum dos estágios históricos da humanidade (caçadores-coletores, agricultores ou funcionários de escritórios). Em vez disso, a maior parcela dos humanos nesse cenário poderia ser de rentistas ociosos que sobrevivessem marginalmente por meio de suas poupanças.[377] Eles seriam muito pobres e obteriam uma pequena renda com o rendimento de suas economias ou subsídios do Estado. Eles viveriam em um mundo com tecnologia extremamente avançada, incluindo não apenas máquinas superinteligentes, mas também medicina antienvelhecimento, realidade virtual, diversas tecnologias de aperfeiçoamento cognitivo e drogas de entretenimento: ainda assim, tudo isso poderia ser, de um modo geral, financeiramente inacessível. Talvez, em vez de usar medicina de aperfeiçoamento, eles usassem drogas que limitassem seu crescimento e diminuíssem seu metabolismo com a intenção de reduzir seu custo de vida (seres que queimassem muitas calorias seriam incapazes de sobreviver com uma renda de subsistência gradualmente decrescente). Conforme nossos números aumentassem e nossa renda média diminuísse ainda mais, poderíamos nos reduzir a alguma estrutura mínima que ainda nos qualificasse para o recebimento de uma pensão — talvez na forma de cérebros minimamente conscientes vivendo em cubas, oxigenados e alimentados por máquinas, lentamente poupando dinheiro suficiente para a reprodução na forma de um clone que seria desenvolvido por um robô técnico.[378]

Frugalidade adicional poderia ser obtida por meio de transferências para um meio digital, já que um substrato para a computação otimizado fisicamente e desenvolvido por uma superinteligência avançada seria mais eficiente do que um cérebro biológico. A migração para o meio digital poderia decair, no entanto, se as emulações fossem consideradas não humanas ou não cidadãs e inelegíveis para receber pensões ou possuir poupanças isentas de cobrança de impostos. Nesse caso, um nicho para os humanos biológicos poderia se manter aberto, ao lado de uma população talvez muito maior de emulações ou inteligências artificiais.

Até agora focamos no destino dos humanos, os quais poderiam ser sustentados por poupanças, subsídios ou salários oriundos de outros humanos que preferissem contratar humanos. Direcionemos nossa atenção agora para algumas das entidades

que até o momento classificamos como “capital”: máquinas que podem ser propriedade de seres humanos, que são construídas e operadas em nome das tarefas funcionais que elas executam e que são capazes de substituir os humanos numa ampla variedade de tarefas. Qual seria a situação dessa força de trabalho nessa nova economia?

Se essas máquinas fossem meramente robôs, simples equipamentos como uma máquina a vapor ou o mecanismo de um relógio, então nenhum comentário adicional seria necessário: haveria uma grande quantidade desse tipo de capital numa economia pós-transição, mas ninguém se importaria com o destino de meros equipamentos inconscientes. Entretanto, caso essas máquinas venham a ter mentes conscientes — se forem construídas de modo que sua operação esteja associada a experiências conscientes (ou se, por alguma outra razão, for dado a elas um status moral) —, então será importante considerar de que maneira o resultado final afetaria essas mentes de máquina. O bem-estar desses trabalhadores com mentes de máquina poderia ser até o aspecto mais importante, uma vez que eles podem vir a ser numericamente dominantes.

Escravidão voluntária, morte casual

Uma questão inicial importante é se essas mentes de máquina serão controladas como capital (escravas) ou se serão contratadas como trabalhadores assalariados livres. No entanto, numa avaliação mais detalhada, é duvidoso que essa questão seja realmente relevante. Há duas razões para isso. Primeiro, se um trabalhador livre num estado malthusiano recebe um salário num nível de subsistência, ele não terá renda disponível depois que pagar por alimento e outras necessidades. Se, em vez disso, o trabalhador for um escravo, seu proprietário pagará por sua subsistência e ele também não terá renda disponível. Em qualquer um dos casos, o trabalhador não adquire nada além de suas necessidades. Segundo, suponha que o trabalhador livre estivesse de alguma forma na posição de receber uma renda acima do nível de subsistência (talvez em virtude de regulamentações favoráveis). Como ele gastaria esse excedente? Investidores poderiam achar mais lucrativo criar trabalhadores que seriam “escravos voluntários” — os quais trabalhariam voluntariamente em troca de salários de subsistência. Os investidores poderiam criar tais trabalhadores copiando trabalhadores que sejam obedientes. Com uma seleção apropriada (e talvez alguma modificação no código), esses investidores seriam capazes de criar trabalhadores que não apenas preferissem oferecer seu trabalho voluntariamente, mas que também escolhessem doar a seus proprietários qualquer renda excedente que recebessem. Oferecer dinheiro ao trabalhador seria então uma forma indireta de fornecer dinheiro ao proprietário ou empregador, mesmo que o trabalhador fosse um agente livre com todos os direitos legais.

Talvez haja objeções em relação ao cenário mencionado, levando em conta que seria difícil projetar uma máquina que quisesse se voluntariar para executar qualquer trabalho a ela designado ou que desejasse doar seus salários para o proprietário. Emulações, em particular, poderiam ser imaginadas como detentoras de desejos mais tipicamente humanos. Mas perceba que, embora o problema do controle original seja difícil, estamos considerando aqui uma condição que viria *após* a transição, uma época em que métodos para a seleção de motivação já

teriam sido presumivelmente aperfeiçoados. No caso das emulações, isso seria possível simplesmente por meio da *seleção* de características pré-existentes no conjunto geral da população humana; e já descrevemos aqui uma série de outros métodos de seleção de motivação. O problema do controle também pode, de algum modo, ser simplificado pela suposição atual de que novas inteligências de máquina ingressariam em uma matriz socioeconômica estável, que já estaria povoada com outros agentes superinteligentes cumpridores das leis.

Consideremos então a difícil situação das máquinas da classe trabalhadora, estejam elas operando como escravas ou agentes livres. Focaremos primeiro nas emulações, o caso mais fácil de se imaginar.

São necessários em torno de quinze a trinta anos para posicionar um novo trabalhador biológico no mercado de trabalho, dependendo do grau exigido de especialização e experiência. Durante esse período, essa nova pessoa precisa ser alimentada, abrigada, cuidada e educada — a um alto custo. Criar uma nova cópia de um trabalhador digital, ao contrário, é tão simples quanto carregar um novo programa na memória do computador. A vida se tornaria, então, barata. Uma empresa poderia adaptar continuamente sua força de trabalho para suprir suas demandas apenas produzindo novas cópias de trabalhadores — e eliminando as cópias que não fossem mais necessárias para liberar recursos computacionais. Isso poderia levar a uma taxa de mortalidade extremamente alta entre os trabalhadores digitais. Muitos poderiam viver apenas o equivalente a um dia.

Existem outras razões além das flutuações de demanda que poderiam motivar empregadores ou donos de emulações a frequentemente “matar” ou “finalizar” seus trabalhadores.[379] Se uma mente emulada, assim como uma mente biológica, demandasse períodos de descanso e sono para funcionar, poderia ser mais barato apagar uma emulação cansada no fim do dia e substituí-la por uma emulação nova e descansada. Como esse procedimento causaria uma amnésia retrógrada, apagando tudo o que tenha sido aprendido durante aquele dia, emulações responsáveis por tarefas que exigissem uma linha longa de pensamento seriam poupadas de apagamentos tão frequentes. Seria difícil, por exemplo, escrever um livro se toda a manhã, quando você se sentasse na sua escrivaninha, você não tivesse memória alguma a respeito do que tivesse produzido nos dias anteriores. Mas outros trabalhos poderiam ser executados adequadamente por agentes que fossem frequentemente reciclados: um vendedor de loja ou um agente de atendimento ao cliente, uma vez treinado, pode precisar se lembrar de novas informações apenas pelo período de vinte minutos.

Uma vez que a reciclagem de emulações impediria a formação de memória e habilidades, algumas emulações poderiam ser colocadas em um tipo de curso intensivo de aprendizagem durante o qual elas rodariam constantemente, incluindo períodos de sono e descanso, mesmo em trabalhos que não demandem necessariamente longas linhas de pensamento. Por exemplo, alguns agentes de atendimento ao cliente poderiam rodar por muitos anos em ambientes otimizados de aprendizagem, auxiliados por treinadores e avaliadores de desempenho. Os que se saíssem melhor poderiam ser utilizados como reprodutores, servindo de modelos para as milhões de novas cópias que seriam criadas diariamente. Muito esforço

seria empregado no aprimoramento do desempenho desses trabalhadores que serviriam como modelos porque até mesmo um pequeno incremento em sua produtividade teria um grande valor econômico quando aplicado em milhões de cópias.

Paralelamente aos esforços para treinamento de trabalhadores-modelo destinados a tarefas específicas, esforços intensos também seriam aplicados na melhoria das tecnologias de emulação como um todo. Avanços nessa área seriam ainda mais valiosos do que avanços feitos com trabalhadores-modelo individuais, uma vez que melhorias tecnológicas gerais poderiam ser usadas em todos os trabalhadores emulados (e potencialmente também em emulações que não sejam de trabalhadores), e não apenas naqueles que tenham uma ocupação em particular. Muitos recursos seriam dedicados à procura de um atalho computacional que permitisse implementações mais eficientes de emulações já existentes, bem como o desenvolvimento de IAs neuromórficas e inteiramente sintéticas. Essa pesquisa seria feita, provavelmente, por emulações executadas em hardwares muito velozes. Dependendo do preço dos recursos computacionais, milhões, bilhões ou trilhões de emulações das mentes humanas mais brilhantes para pesquisa (ou versões melhoradas) poderiam trabalhar 24 horas por dia em busca de avanços nas fronteiras do conhecimento da inteligência de máquina; e algumas delas poderiam operar algumas ordens de magnitude mais rápido do que cérebros biológicos.[380] Essa é uma boa razão para considerarmos que a era das emulações humanas será breve — um interlúdio *muito* breve em tempo sideral — e que logo dará lugar a uma era de inteligência artificial imensamente superior.

Já vimos diversas razões que podem levar os empregadores de emulações trabalhadoras a reduzir os seus rebanhos periodicamente: flutuações na demanda por diferentes tipos de trabalhadores, diminuição de custos com a eliminação de emulação de sono e descanso, além da introdução de novos e melhores modelos. Preocupações em relação à segurança também podem fornecer outras razões. Para prevenir que trabalhadores elaborem algum tipo de conspiração ou plano subversivo, emulações que ocupem posições centrais podem ser programadas para rodar apenas por um breve período de tempo, sendo restauradas constantemente para um estado predefinido e armazenado anteriormente.[381]

Esses estados predefinidos para os quais as emulações seriam restauradas seriam cuidadosamente preparados e avaliados. Uma típica emulação de vida curta poderia acordar em um estado mental descansado que seja otimizado para a lealdade e produtividade. Ela se lembraria de ter se formado como primeira colocada na sua turma depois de (muitos) anos subjetivos de intenso treinamento, de ter aproveitado um fim de semana revigorante e uma boa noite de sono, de ter assistido a uma palestra motivacional empolgante e de ter escutado uma música agitada e, naquele momento, ela estaria ansiosa para finalmente voltar ao trabalho e fazer o máximo pelo seu empregador. Sua cabeça não estaria repleta de pensamentos a respeito de sua morte iminente ao término de um dia de trabalho. Emulações com neuroses a respeito da morte ou outros problemas do tipo seriam menos produtivas e não teriam sido selecionadas.[382]

Trabalhar da forma mais eficiente possível seria divertido?

Uma variável importante na avaliação do quanto uma condição hipotética como essa seria desejável é o estado hedônico de uma emulação típica.[383] Essa dura jornada de trabalho resultaria em sofrimento ou felicidade para um típico trabalhador emulado?

Devemos resistir à tentação de projetar nossos próprios sentimentos nesse trabalhador emulado imaginário. A questão aqui não é se *você* se sentiria feliz se tivesse que trabalhar constantemente e nunca mais pudesse se encontrar com seus entes queridos — um destino terrível, muitos concordariam.

É moderadamente mais relevante considerar qual é a experiência hedônica que um humano comum experimenta atualmente durante suas horas de trabalho. Estudos feitos com indivíduos do mundo todo a fim de avaliar o quão felizes eles se consideravam demonstram que a maioria dos entrevistados se considerava "feliz" ou "muito feliz" (3,1 em uma escala de 1 a 4).[384] Estudos sobre o sentimento médio da população, os quais questionaram os participantes a respeito da frequência com que eles experimentaram diversos estados mentais positivos ou negativos recentemente, tendem a gerar resultados similares (com um efeito de satisfação geral de cerca de 0,52 em uma escala de -1 a 1). Existe um pequeno efeito positivo da renda per capita de um país no nível médio de bem-estar.[385] Entretanto, é perigoso extrapolar esses resultados para o estado hedônico de futuros trabalhadores emulados. Uma razão para tanto é que sua condição seria muito diferente: por um lado, eles poderiam trabalhar muito mais; por outro, eles estariam livres de doenças, dores, fome, odores ruins e assim por diante. No entanto, tais considerações não dão conta de responder nossa pergunta. A questão mais importante aqui é que o teor hedônico seria fácil de ser ajustado por meio de um equivalente digital a uma droga ou neurocirurgia. Isso significa que seria um erro inferir o estado hedônico de futuras emulações a partir das condições externas de suas vidas apenas imaginando como nós mesmos e outras pessoas como nós iríamos nos sentir nas mesmas circunstâncias. O estado hedônico seria uma questão de escolha. Dentro do modelo que estamos considerando agora, a escolha seria feita pelos detentores de capital que estivessem buscando maximizar os retornos dos seus investimentos em trabalhadores emulados. Consequentemente, a questão a respeito do quanto as emulações seriam felizes se resume à questão em torno de quais estados hedônicos seriam mais produtivos (nos diversos tipos de tarefas nas quais as emulações seriam empregadas).

Aqui, novamente, podemos tentar delinear uma inferência a partir de observações sobre a felicidade humana. Se é verdade que, ao longo da história e na maioria dos lugares e ocupações, as pessoas foram, de modo geral, ao menos razoavelmente felizes, isso criaria alguma presunção de que o mesmo ocorreria em um cenário de pós-transição como o que estamos considerando. Para ser claro, o argumento nesse caso não seria o de que mentes humanas possuem uma predisposição à felicidade de forma que encontrariam satisfação nessas novas condições, mas que um certo nível padrão de felicidade se mostrou adaptativo para as mentes humanas no passado, de forma que talvez um nível similar de felicidade se mostrará

adaptativo para mentes semelhantes à humana no futuro. Mesmo assim, essa formulação revela um ponto fraco de tal inferência: as disposições mentais que foram adaptativas aos hominídeos caçadores e coletores que vagavam pelas savanas africanas podem não ser necessariamente adaptativas para emulações modificadas que vivam dentro de realidades virtuais num período pós-transição. Podemos certamente ter *esperança* de que os futuros trabalhadores emulados sejam tão ou mais felizes do que foram os trabalhadores humanos comuns ao longo da história, mas ainda não vimos nenhuma razão convincente para supor que isso de fato aconteceria (dentro do cenário multipolar de livre mercado que estamos examinando no momento).

Considere a possibilidade de que a razão pela qual a felicidade prevalece entre os humanos (para qualquer grau limitado em que ela prevaleça) seja que o estado de espírito alegre possui uma função de sinalização no ambiente de adaptação evolutiva. Passar aos outros membros do grupo social a impressão de estar em condições prósperas — com boa saúde, com boa reputação entre seus pares e confiantes na continuidade dessas boas condições — pode ter aumentado a popularidade dos indivíduos que possuíam tal atitude. Uma tendência à alegria pode ter sido evolutivamente selecionada, o que fez com que os processos químicos neurais nos humanos se voltassem em direção a sentimentos positivos comparados com o que poderia ter sido mais eficiente de acordo com critérios materialistas mais simples. Se esse fosse o caso, então o futuro da *joie de vivre* pode depender da manutenção inalterada da alegria como função de sinalização social num mundo pós-transição: questão para a qual voltaremos em breve.

Mas e se almas felizes dissiparem mais energia que as almas infelizes? Provavelmente, os indivíduos felizes seriam mais propensos a surtos criativos e voos da imaginação — comportamento que futuros empregadores podem desvalorizar na maioria de seus funcionários. Talvez uma obsessão ansiosa e taciturna por simplesmente concluir o trabalho sem cometer erros será a atitude maximizadora da produtividade na maioria dos ramos profissionais. Não estamos afirmando aqui que esse seria o caso, mas apenas que não sabemos exatamente qual será. Contudo, devemos considerar o quão desagradável seria caso se mostrasse verdadeira uma hipótese pessimista como essa a respeito de um futuro estado malthusiano: não apenas em virtude do custo de oportunidade gerado pela falha na criação de algo melhor — que seria enorme —, mas também porque essa situação poderia ser ruim em si mesma, possivelmente muito pior do que o estado malthusiano original.

Nós raramente nos esforçamos ao máximo. Quando o fazemos, esse esforço é às vezes algo doloroso. Imagine ter que ficar correndo em uma esteira inclinada — batimentos cardíacos acelerados, dores musculares, falta de ar. Uma olhada no cronômetro: a sua próxima pausa, que também será a sua morte, se dará em 49 anos, três meses, vinte dias, quatro horas, 56 minutos e doze segundos. Você desejaria não ter nascido.

Novamente, não estamos afirmando que isso acontecerá dessa forma, mas não temos certeza de que não acontecerá. Poderíamos certamente defender um cenário mais otimista. Por exemplo, não há uma razão óbvia pela qual as emulações

precisariam sofrer lesões corporais ou adoecerem: a eliminação do desgaste físico seria uma grande melhoria em relação ao estágio em que nos encontramos agora. Além disso, uma vez que a realidade virtual tem um custo de criação relativamente baixo, as emulações poderiam trabalhar em lugares suntuosos — em esplêndidos palácios no topo de montanhas, em terraços com vista para um bosque, ou na beira de uma lagoa azulada —, apenas com alguns ajustes na iluminação, temperatura, cenário e decoração; elas estariam livres de fumaça, barulho, rajadas de vento e zunido de insetos; estariam vestidas com roupas confortáveis, se sentiriam limpas, focadas e bem alimentadas. E mais importante: se — e isso parece perfeitamente possível — o estado mental humano ideal para uma maior produtividade na maioria das categorias profissionais se definir pelo entusiasmo e alegria, então a era da economia de emulações poderia ser bastante aprazível.

Existiria, de qualquer maneira, uma grande vantagem em organizar as coisas de modo que algo ou alguém pudesse intervir e fazer correções caso a trajetória padrão, por acaso, seguisse em direção a uma distopia. Também poderia ser desejável que houvesse algum plano de fuga que permitisse a escolha pela morte e esquecimento caso a qualidade de vida caísse a um nível em que a aniquilação se tornasse preferível à continuação da existência.

Inconscientes terceirizados?

No longo prazo, conforme a era das emulações ceder lugar à era da inteligência artificial (ou se a inteligência de máquina for alcançada diretamente pela via da IA sem passar pelo estágio da emulação cerebral), é possível que o prazer e a dor desapareçam completamente em um cenário multipolar, uma vez que um mecanismo de recompensa hedônico pode não ser o sistema de motivação mais eficiente para um agente artificial complexo (um que, ao contrário da mente humana, não seja sobrecarregado com um legado de *wetware* animal). Talvez um sistema de motivação mais avançado pudesse ser baseado em uma representação explícita de uma função de utilidade ou em alguma outra arquitetura que não tenha análogos funcionais exatos ao prazer e à dor.

Um cenário multipolar semelhante, embora ligeiramente mais radical — que poderia envolver a eliminação de quase todo o valor do futuro —, é aquele em que o proletariado não seria sequer consciente. Essa possibilidade é mais provável de acontecer num cenário com IAs, que podem ser estruturadas de formas bem distintas da inteligência humana. Mas mesmo que a inteligência de máquina fosse alcançada inicialmente através da emulação completa do cérebro, resultando em mentes digitais conscientes, as forças competitivas desencadeadas em uma economia pós-transição poderiam facilmente levar ao surgimento de formas de inteligência de máquina cada vez menos neuromórficas, tanto porque a IA sintética seria criada do zero como porque as emulações poderiam, através de sucessivas modificações e melhorias, se afastar progressivamente de sua forma humana original.

Considere um cenário no qual, após o desenvolvimento da tecnologia de emulação, o progresso contínuo em neurociência e ciência da computação (acelerados pela presença de mentes digitais que serviriam tanto como pesquisadores quanto como cobaias para testes) tornasse possível isolar módulos

cognitivos individuais em uma emulação e ligá-los a módulos isolados de outras emulações. Um período de treinamento e ajustes poderia ser necessário antes que módulos diferentes pudessem colaborar efetivamente; mas módulos construídos de acordo com padrões comuns poderiam se ligar mais rapidamente a outros módulos com o mesmo padrão. Isso tornaria os módulos padronizados mais produtivos, criando uma pressão em direção a uma maior padronização.

As emulações poderiam, então, começar a terceirizar uma crescente parcela de suas funcionalidades. Para que aprender aritmética quando você pode mandar suas questões matemáticas para a Módulos-Gauss, S.A.? Para que ser articulado se você pode contratar a Coleridge Conversas para transformar seus pensamentos em palavras? Por que tomar decisões a respeito de sua vida pessoal quando existem módulos executivos que podem fazer uma varredura em seu sistema de objetivos e gerenciar recursos para alcançar seus objetivos de uma maneira melhor do que você conseguiria se tivesse feito sozinho? Algumas emulações podem preferir manter a maioria de suas funcionalidades e lidar com as tarefas elas mesmas em vez de delegá-las a terceiros que poderiam concluí-las de maneira mais eficiente. Essas emulações seriam como indivíduos que possuem um hobby, que gostam de plantar suas próprias verduras ou tricotar seus próprios casacos. Tais emulações amadoras seriam menos eficientes; e se existisse um fluxo líquido de recursos que partisse dos participantes da economia menos eficientes em direção aos mais eficientes, essas emulações amadoras acabariam perdendo.

Os cubos de tempero dos intelectos discretos mais similares aos humanos se dissolveriam, então, em uma sopa algorítmica.

É concebível que a eficiência máxima seria obtida através do agrupamento de capacidades em agregados que correspondam aproximadamente à arquitetura cognitiva de uma mente humana. Talvez fosse o caso, por exemplo, de adaptar o módulo matemático ao módulo de linguagem, e então ambos tivessem que ser adaptados ao módulo executivo, de maneira que os três passassem a trabalhar em conjunto. A terceirização cognitiva seria, dessa forma, quase que totalmente impraticável. Mas na ausência de qualquer razão convincente para confiar que algo do tipo poderia acontecer, devemos contemplar a possibilidade de que arquiteturas cognitivas semelhantes à da mente humana sejam ótimas apenas se levarmos em conta as restrições da neurologia humana (ou em hipótese nenhuma). Quando a construção de arquiteturas que não possam ser bem implementadas em redes neurais biológicas se tornar possível, surgirá também espaço para novos projetos; e os ótimos globais nesse espaço estendido não precisarão ser semelhantes a tipos familiares de mentalidade. Desse modo, organizações cognitivas similares à humana não teriam um nicho em economias ou ecossistemas competitivos no pós-transição.[386]

Talvez houvesse nichos para agrupamentos menos complexos (como os módulos individuais) ou mais complexos (como grandes agrupamentos de módulos), ou ainda aqueles com complexidade similar à da mente humana, mas com arquiteturas radicalmente diferentes. Esses agrupamentos teriam algum valor intrínseco? Deveríamos dar boas-vindas a um mundo no qual tais agrupamentos estranhos teriam substituído os agrupamentos humanos?

A resposta pode depender da natureza específica desses agrupamentos estranhos. O mundo possui atualmente diversos níveis de organização. Algumas entidades altamente complexas, tais como as corporações multinacionais e os Estados-nação, têm seres humanos como seus constituintes e mesmo assim nós atribuímos a esses agrupamentos com alto grau de complexidade apenas um valor instrumental. Estados e corporações não possuem (imagina-se) consciência, além da consciência das pessoas que os formam: eles não podem sentir dor ou prazer fenomenais ou experimentar qualquer espécie de *qualia*. Nós os valorizamos na medida em que eles servem às necessidades humanas e, quando deixam de fazê-lo, os "matamos" sem dó nem piedade. Existem também as entidades de baixa complexidade e, geralmente, a essas também é negado um status moral. Nós não vemos mal algum em apagar um aplicativo de nossos smartphones e não acreditamos que uma neurocirurgiã esteja fazendo nada de errado quando retira um módulo defeituoso do cérebro de um epilético. Em relação aos agrupamentos que possuem uma organização exótica e um nível de complexidade similar ao do cérebro humano, a maioria de nós talvez pudesse considerá-los dotados de alguma importância moral apenas se, ao nosso ver, eles demonstrassem capacidade ou potencial para a experiência consciente.[387]

Desse modo, podemos imaginar, como um caso extremo, uma sociedade altamente avançada tecnologicamente, contendo diversas estruturas complexas, algumas delas muito mais intrincadas e inteligentes do que qualquer uma que exista no planeta atualmente — uma sociedade que, todavia, não possua nenhum tipo de ser consciente ou cujo bem-estar tenha qualquer significado moral. De certa forma, essa seria uma sociedade despovoada. Seria uma sociedade de milagres econômicos e assombros tecnológicos, mas ninguém para se beneficiar deles. Uma Disneylândia sem crianças.

Evolução não é necessariamente para cima

A palavra "evolução" é, em geral, usada como sinônimo de "progresso", talvez refletindo uma imagem acrítica da evolução como uma força para o bem. A fé descabida em benefícios inerentes ao processo evolucionário pode prejudicar uma avaliação justa do quão desejável seria um cenário multipolar no qual o futuro da vida inteligente fosse determinado por dinâmicas competitivas. Qualquer avaliação desse tipo precisa se apoiar em alguma opinião (ao menos implícita) a respeito da distribuição de probabilidade de diferentes fenótipos que venham a ser bem adaptados em uma sopa vital digital no pós-transição. Seria difícil, mesmo na melhor das circunstâncias, extrair uma resposta clara e exata da inevitável mistura de incertezas que permeiam essas questões: ainda mais se adicionarmos uma camada de lama panglossiana.

Uma possível fonte de fé na evolução livre é a trajetória aparentemente ascendente que o processo evolucionário exibiu no passado. Iniciando com organismos replicadores rudimentares, a evolução foi capaz de produzir organismos cada vez mais "avançados", incluindo criaturas com mentes, consciência, linguagem e razão. Mais recentemente, processos culturais e tecnológicos, que apresentam uma vaga similaridade com a evolução biológica, permitiram que os humanos se

desenvolvessem em um ritmo acelerado. Em uma escala de tempo tanto geológica como histórica, o quadro geral parece mostrar uma tendência global que aponta para um sempre crescente grau de complexidade, conhecimento, consciência e uma organização coordenada com o intuito de alcançar objetivos: uma tendência que, sejamos francos, poderia ser chamada de “progresso”.[388]

A imagem da evolução como um processo que produz efeitos benignos de forma consistente é algo difícil de ser conciliado com o enorme sofrimento que percebemos tanto em humanos quanto no mundo natural. Aqueles que valorizam os feitos da evolução talvez o façam de um ponto de vista mais estético do que ético. Entretanto, a questão pertinente aqui não se trata do tipo de futuro que seria mais fascinante ver representado em um livro de ficção científica ou num documentário sobre a natureza, mas que tipo de futuro seria bom vivenciar: duas questões bastante distintas.

Além do mais, não temos nenhuma razão para pensar que qualquer progresso que por ventura ocorreu tenha sido inevitável. Grande parte dele pode ter sido pura sorte. Essa objeção é baseada no fato de que um efeito de seleção por observação filtra a evidência que podemos ter a respeito do sucesso de nosso próprio desenvolvimento evolutivo.[389] Suponha que em 99,9999% de todos os planetas onde a vida surgiu, ela tenha se extinguido antes do ponto em que seus observadores inteligentes pudessem começar a se questionar a respeito de suas origens. O que poderíamos esperar observar se esse fosse o caso? É possível argumentar que deveríamos ter a expectativa de observar algo semelhante ao que de fato nós observamos. A hipótese de que sejam pequenas as chances de evolução de vida inteligente em um dado planeta não serve para prever que de fato nos encontraríamos em um planeta onde a vida se extinguiu em estágios iniciais; ao contrário, isso pode ser útil para prever que deveríamos estar em um planeta no qual a vida inteligente evoluiu, mesmo se tais planetas constituíssem uma pequena fração de todos os planetas onde a vida primitiva apresentou alguma evolução. Portanto, o longo histórico de registros de vida na Terra oferece pouco suporte à afirmação de que havia uma alta probabilidade em relação ao surgimento de organismos complexos em nosso planeta — e menos ainda que isso tenha sido inevitável.[390]

Em terceiro lugar, mesmo que as condições atuais fossem perfeitas ou se tivéssemos a capacidade de provar que evoluímos inevitavelmente a partir de algum estado primordial genérico, não existiria nenhuma garantia de que essa tendência de melhoria continuaria indefinidamente no futuro. Essa afirmação pode ser sustentada ainda que eliminemos a possibilidade de um evento cataclísmico capaz de causar uma extinção em massa e até mesmo se considerarmos que os processos evolutivos continuarão a produzir sistemas de complexidade crescente.

Sugerimos anteriormente que trabalhadores com uma inteligência de máquina que fossem selecionados para apresentar produtividade máxima trabalhariam arduamente e observamos que não sabemos o quanto eles seriam felizes. Também aventamos a possibilidade de que as formas de vida mais adaptáveis a uma futura e competitiva sopa vital digital poderiam nem ao menos ser conscientes. Além da completa inexistência de prazer ou consciência, também poderiam estar ausentes

diversas outras qualidades que muitos julgariam indispensáveis a uma vida boa. Os seres humanos valorizam a música, o humor, o romance, a arte, a diversão, a dança, conversas, filosofia, literatura, aventura, descobertas, comidas e bebidas, amizade, criação de filhos, esportes, natureza, tradições e espiritualidade, entre muitas outras coisas. Não há garantia de que qualquer uma dessas coisas se manteriam adaptativas. Talvez a adaptação máxima ocorra especificamente por meio do trabalho intenso e ininterrupto, de natureza monótona e repetitiva, destituído de qualquer característica lúdica e objetivando apenas o crescimento da oitava casa decimal de algum indicador econômico qualquer. Os fenótipos selecionados teriam, então, vidas sem nenhuma das qualidades mencionadas acima e, dependendo da axiologia empregada, o resultado poderia parecer abominável, inútil ou apenas empobrecido, e de qualquer maneira estaria muito longe de uma utopia que merecesse o entusiasmo de alguém.

Seria possível nos perguntar como um quadro tão sombrio poderia ser consistente com o fato de que nós, atualmente, apreciamos música, humor, romance, arte etc. Se esses comportamentos são realmente tão “supérfluos”, então o que explica que tenham sido tolerados e até mesmo promovidos pelos processos evolutivos que moldaram nossa espécie? O fato de que o homem moderno se encontra atualmente em um desequilíbrio evolutivo não poderia justificar, já que nossos ancestrais no Plistoceno também perdiam tempo com tais distrações. Muitos dos comportamentos em questão não são nem mesmo exclusivos ao *Homo sapiens*. Exibições extravagantes são encontradas em uma ampla variedade de contextos, desde a seleção sexual no reino animal até as lutas por prestígio político entre os Estados-nação.[391]

Embora a completa explicação evolucionária para cada um desses comportamentos esteja além do escopo deste livro, é possível perceber que alguns deles servem a funções que podem não ser tão relevantes em um contexto de inteligência de máquina. Brincadeiras, por exemplo, que ocorrem apenas em algumas espécies, principalmente entre os mais jovens, são, sobretudo, uma maneira pela qual um animal mais jovem aprende habilidades de que precisará mais tarde durante a vida. Quando emulações puderem ser criadas já em idade adulta e com um repertório completo de habilidades, ou quando o conhecimento e as técnicas adquiridas por uma IA puderem ser transferidas diretamente para outra, a necessidade de comportamentos lúdicos pode se tornar menos difundida.

Diversos outros exemplos de comportamentos humanísticos podem ter evoluído como sinais, difíceis de ser imitados, de qualidades que não podemos observar facilmente de forma direta, como a resiliência mental e corporal, o status social, a qualidade dos aliados, a habilidade ou vontade de vencer uma luta ou a posse de recursos. O rabo do pavão é um exemplo clássico: apenas os pavões melhor adaptados conseguem exibir uma plumagem realmente extravagante, e as pavoas evoluíram para achá-las atrativas. Tanto quanto as características morfológicas, as características comportamentais também podem sinalizar uma adaptação genética ou outros atributos socialmente relevantes.[392]

Como exibições extravagantes são comuns tanto entre os humanos quanto em outras espécies, podemos nos perguntar se elas também não fariam parte do

repertório de formas de vida mais tecnologicamente avançadas. Mesmo que houvesse pouco uso instrumental para brincadeiras e musicalidade, ou mesmo para a consciência, em uma futura ecologia de processamento inteligente de informação, essas características, ainda assim, não poderiam conferir alguma vantagem evolutiva a quem as possuísse, uma vez que são sinais confiáveis de outras qualidades adaptativas?

Embora a possibilidade de uma harmonia preestabelecida entre o que é valioso para nós e o que seria adaptativo em uma futura ecologia digital seja algo difícil de ser excluído, existem razões para o ceticismo. Considere, primeiramente, que muitas das exibições mais suntuosas que observamos na natureza estão ligadas à seleção sexual.[393] Já a reprodução entre formas de vida tecnologicamente maduras pode ser, em contraste, predominantemente ou exclusivamente assexuada.

Em segundo lugar, agentes tecnológicos avançados podem vir a ter maneiras novas e confiáveis de comunicar informações a seu respeito, as quais não dependeriam de exibições suntuosas. Mesmo hoje em dia, quando bancos querem conferir se alguém possui um bom histórico de crédito, essas instituições tendem a confiar mais em evidências documentais, como posse de imóveis e extratos bancários, do que em exibicionismo e ostentação, como roupas caras ou um Rolex. No futuro, talvez seja possível contratar firmas de auditoria para verificar detalhadamente o histórico de comportamento de um agente digital, efetuar testes em ambientes simulados ou inspecionar diretamente o código-fonte para verificar se um agente cliente apresenta um atributo que ele afirme possuir. Concordar em sinalizar qualidades por meio de uma auditoria desse tipo poderia ser mais eficiente do que sinalizá-las através de exibições extravagantes. O custo de se *falsificar* uma sinalização profissionalmente mediada como essa seria alto — principal motivo pelo qual essa sinalização seria confiável —, mas o custo para comunicá-la seria muito inferior quando *verdadeira* do que seria para transmitir um sinal equivalente de maneira extravagante.

Em terceiro lugar, nem todas essas exibições suntuosas apresentam um valor intrínseco ou são socialmente desejáveis. Muitas delas são simplesmente um desperdício. Exibições extravagantes como as cerimônias *potlatch* dos Kwakiutl, uma forma de competição por status entre chefes rivais, envolviam a destruição pública de grandes quantidades de riqueza acumulada.[394] Arranha-céus, megaiates e foguetes lunares podem ser vistos como equivalentes contemporâneos. Apesar de ser possível afirmar plausivelmente que atividades como música e humor melhorem de forma intrínseca a qualidade de vida humana, é duvidoso que possamos fazer uma afirmação semelhante em relação à busca custosa por acessórios de moda ou outros símbolos consumistas de status. Pior do que isso, exibições extravagantes podem indiscutivelmente causar danos, como uma atitude de “macho” que leva à violência entre gangues ou bravatas militares. Desse modo, mesmo se formas futuras de vida inteligente se utilizassem de uma sinalização suntuosa, ainda assim seria uma questão em aberto se o sinal seria valioso de alguma forma — se ele seria como o canto melodioso de um rouxinol ou como o coaxar monossilábico de um sapo (ou o latido incessante de um cachorro).

Formação de um *singleton* no pós-transição?

Mesmo se o resultado imediato da transição para a inteligência de máquina fosse multipolar, ainda existiria a possibilidade de que um *singleton* viesse a se desenvolver posteriormente. Tal desenvolvimento daria continuidade a uma aparente tendência de longo prazo em direção à integração política em escalas maiores, levando-o à sua conclusão natural.[395] Como isso poderia ocorrer?

Uma segunda transição

Um dos caminhos pelo qual um resultado inicial multipolar poderia convergir em um *singleton* no pós-transição seria aquele onde existisse, depois da transição inicial, uma segunda transição tecnológica grande e rápida o suficiente para conceder uma vantagem estratégica decisiva a alguma das potências remanescentes: uma dessas potências poderia se valer dessa oportunidade para estabelecer um *singleton*. Essa segunda transição hipotética poderia ser ocasionada pela descoberta de alguma superinteligência de nível mais avançado.[396] Por exemplo, se a primeira leva de máquinas superinteligentes for baseada em emulações, então uma segunda geração poderia surgir a partir do momento em que as emulações fossem bem-sucedidas nas pesquisas pelo desenvolvimento de uma IA capaz de se aperfeiçoar de forma autônoma. (Alternativamente, uma segunda transição poderia ser desencadeada por uma descoberta no campo da nanotecnologia ou alguma outra tecnologia de caráter geral ou militar que ainda não foi concebida.)

O ritmo de desenvolvimento após a transição inicial seria extremamente rápido. Possivelmente até mesmo um pequeno intervalo entre a potência dominante e seu competidor mais próximo poderia, dessa forma, resultar em uma vantagem estratégica decisiva para a potência dominante durante a segunda transição. Suponha, por exemplo, que dois projetos iniciem a primeira transição apenas com alguns dias de intervalo entre si, e que a partida seja lenta o suficiente para que esse intervalo não dê ao projeto líder uma vantagem estratégica decisiva em nenhum momento ao longo da partida. Ambos os projetos emergiriam como potências superinteligentes, embora um deles permaneceria alguns dias à frente do outro. Mas os avanços passariam a ocorrer em uma escala de tempo compatível com pesquisas conduzidas por uma superinteligência de máquina — talvez milhares ou milhões de vezes mais rápidas do que as pesquisas conduzidas em uma escala de tempo biológica humana. Sendo assim, o desenvolvimento de tecnologias para a segunda transição poderia ocorrer em questão de dias, horas ou minutos. Mesmo que o projeto líder tivesse apenas alguns meros dias de vantagem, uma nova descoberta poderia lançá-lo a uma vantagem estratégica decisiva. Note, contudo, que se a difusão tecnológica (via espionagem ou algum outro meio) acelerasse no mesmo ritmo que os desenvolvimentos tecnológicos, então esse efeito poderia ser anulado. O que permaneceria relevante seria o quão acentuada a segunda transição pudesse se tornar, ou seja, a velocidade na qual ela ocorreria em relação à velocidade geral dos eventos no período posterior à pós-transição. (Nesse sentido, quanto mais rapidamente os eventos ocorressem no pós-transição, menos acentuada a segunda transição tenderia a ser.)

Também é possível especular que uma vantagem estratégica decisiva teria maior probabilidade de ser usada para formar um *singleton* se ela ocorresse durante a segunda transição (ou numa transição subsequente). Após a primeira transição, as lideranças seriam elas mesmas superinteligentes ou teriam acesso aos conselhos de uma superinteligência, o que evidenciaria as implicações das opções estratégicas disponíveis. Além disso, após a primeira transição, é possível que ocorram movimentos preventivos contra potenciais competidores que poderiam ser menos perigosos para o agressor. Se as mentes que tiverem papel de liderança após a primeira transição forem digitais, elas poderiam ser copiadas e, assim, estariam menos vulneráveis a contra-ataques. Mesmo que o defensor tivesse a capacidade de matar nove décimos da população do agressor em um ataque retaliatório, isso dificilmente o deteria se os mortos pudessem ser imediatamente ressuscitados a partir de backups redundantes. A devastação de infraestrutura (que pode ser reconstruída) também poderia ser tolerada por mentes digitais com expectativas de vida praticamente ilimitadas e que estivessem planejando maximizar seus recursos e influência em uma escala de tempo cosmológica.

Superorganismos e economias de escala

O tamanho de agrupamentos humanos coordenados, como empresas e nações, é influenciado por diversos parâmetros — tecnológicos, militares, financeiros e culturais — que podem variar de um período histórico para outro. Uma revolução de inteligência de máquina poderia acarretar mudanças profundas em muitos desses parâmetros. Talvez essas mudanças facilitassem o surgimento de um *singleton*. Embora não seja possível, sem estudar detalhadamente quais seriam essas prováveis mudanças, excluir a possibilidade contrária — de que as mudanças facilitariam uma fragmentação em vez de uma unificação —, podemos notar que o aumento de variação ou incerteza com o qual lidamos aqui é capaz, por si mesmo, de construir uma base que nos daria uma maior credibilidade em relação ao potencial surgimento de um *singleton* do que no caso contrário. Uma revolução de inteligência de máquina poderia, por assim dizer, agitar as coisas — embaralhar as cartas de maneira a tornar possíveis realinhamentos geopolíticos que, de outro modo, pareceriam improváveis.

Uma análise completa de todos os fatores que poderiam influenciar a escala da integração política nos afastaria muito do escopo deste livro: uma revisão de toda a literatura relevante de economia e ciência política preencheria facilmente uma obra inteira. Teremos que nos limitar a fazer breves alusões a alguns fatores, mais precisamente aos aspectos da digitalização de agentes que poderiam facilitar a centralização do controle.

Carl Shulman argumentou que, em uma população de emulações, as pressões causadas pela seleção natural poderiam favorecer o surgimento de "superorganismos", ou seja, grupos de emulações prontas para se sacrificar em prol de seu clã.[397] Os superorganismos não teriam os problemas de agência que atingem organizações cujos membros possuem seus próprios interesses. Assim como as células em nosso corpo, ou animais individuais em colônias de insetos eussociais, as emulações que fossem completamente altruístas em relação às suas

cópias-irmãs cooperariam umas com as outras, mesmo na ausência de elaborados esquemas de incentivo.

Os superorganismos teriam uma vantagem particularmente forte se a deleção não consensual (ou suspensão por tempo indefinido) de emulações individuais fosse proibida. Firmas ou países que empregassem emulações e insistissem em sua autopreservação ficariam sobrecarregados com um compromisso infindável de pagar pela manutenção de trabalhadores obsoletos ou redundantes. Em contrapartida, organizações nas quais as emulações deletassem a si mesmas voluntariamente quando seus serviços não fossem mais necessários poderiam se adaptar mais facilmente às flutuações de demanda. Essas organizações poderiam experimentar livremente, reproduzindo variações de seus trabalhadores e mantendo apenas os mais produtivos.

Se a deleção involuntária *não* for proibida, então a vantagem comparativa de emulações eussociais seria reduzida, mas talvez não totalmente eliminada. É provável que empregadores de emulações adeptas ao autossacrifício ainda pudessem obter alguns ganhos em eficiência em virtude da redução nos problemas de agência na organização como um todo, inclusive sendo poupados do problema de ter que derrotar qualquer resistência que as emulações pudessem impor contra sua própria deleção. De modo geral, os ganhos de produtividade em possuir trabalhadores dispostos a sacrificar a própria vida em prol do bem comum são um caso específico dos benefícios aos quais uma organização tem acesso por contar com membros que sejam tão fanaticamente devotados a ela. Tais membros não iriam apenas pular para dentro de suas próprias covas em nome da organização e trabalhar longas horas em troca de um pequeno salário: eles também se afastariam de disputas políticas e tentariam consistentemente agir de acordo com o que acreditassem ser o melhor para a organização, o que reduziria a necessidade de supervisão e restrições burocráticas.

Se a única maneira de alcançar tal dedicação fosse restringir a adesão apenas a cópias-irmãs (para que todas as emulações em um superorganismo em particular fossem geradas a partir do mesmo modelo), então os superorganismos teriam uma desvantagem, pois ficariam limitados a escolher entre uma gama menor de habilidades em comparação às organizações rivais, desvantagem essa que poderia ou não ser grande o suficiente para superar as vantagens de não precisar lidar com problemas de agência internos.[398] Essa desvantagem seria imensamente amenizada se um superorganismo pudesse ao menos possuir membros com diferentes tipos de treinamento. Nesse caso, mesmo que todos os seus membros tenham sido originados a partir de um mesmo modelo, sua força de trabalho ainda poderia contribuir com uma gama diversificada de habilidades. Iniciando com um modelo de emulação com um talento polímato, as futuras linhagens poderiam ser divididas em diferentes programas de treinamento, uma cópia aprenderia contabilidade, outra engenharia elétrica e assim por diante. Isso produziria um grupo com habilidades diversas, mesmo que não seja um grupo com talentos diversos. (Para uma máxima diversidade, talvez fosse necessário o uso de mais de um modelo de emulação.)

A propriedade essencial de um superorganismo não consiste no fato de que ele seja composto por cópias provenientes de um único progenitor, mas que todos os agentes que o compõem estejam totalmente comprometidos com um objetivo comum. A habilidade de criar um superorganismo pode, desse modo, ser encarada como dependente de uma solução parcial para o problema do controle. Enquanto uma solução totalmente geral para o problema do controle capacitaria alguém a criar um agente com um objetivo final arbitrário qualquer, a solução parcial necessária para a criação de um superorganismo demandaria meramente a habilidade de produzir múltiplos agentes com o mesmo objetivo final (para alguns objetivos finais não triviais, mas não necessariamente arbitrários).[399]

A principal consideração levantada nesta subseção não está, portanto, necessariamente limitada aos grupos de emulações monoclonais e é possível abordá-la mais amplamente para evidenciar que ela se aplica a uma gama variada de cenários multipolares de inteligência de máquina. O caso é que certos tipos de avanços em seleção de motivação, que podem se tornar viáveis uma vez que os atores sejam digitais, poderiam ajudar a superar algumas das ineficiências que atualmente causam dificuldades a grandes organizações humanas e são um contrapeso nas economias de escala. Ultrapassados esses limites, as organizações — sejam elas empresas, nações ou outras entidades políticas ou econômicas — poderiam aumentar em tamanho. Esse é um fator que facilitaria o surgimento de um *singleton* no pós-transição.

Uma área na qual superorganismos (ou outros agentes digitais com motivações parcialmente selecionadas) podem se sobressair é na coerção. Um estado pode usar métodos de seleção de motivação para assegurar que sua força policial, militar, serviço de inteligência e administração civil sejam totalmente leais. De acordo com Shulman:

> Estados predefinidos [de alguma emulação leal que já tenha sido cuidadosamente preparada e verificada] poderiam ser copiados bilhões de vezes para formar equipes militares, burocráticas e policiais que sejam ideologicamente uniformes. Após um curto período de trabalho, cada cópia seria substituída por uma nova cópia com o mesmo estado predefinido, prevenindo-se assim desvios ideológicos. Dentro de uma dada jurisdição, essa capacidade permitiria observações e regulamentações extremamente detalhadas: existiria, então, uma dessas cópias para cada cidadão. Isso poderia ser usado para proibir o desenvolvimento de armas de destruição em massa, para impor regulamentações a experimentos ou reprodução de emulações cerebrais, para impor uma constituição liberal e democrática ou para criar um regime totalitário terrível e permanente.[400]

O efeito de primeira ordem de tal capacidade parece ser a consolidação do poder, que ficaria, muito provavelmente, concentrado nas mãos de poucos.

Unificação por tratado

Uma colaboração internacional em um mundo multipolar no pós-transição poderia trazer ganhos potencialmente grandes. Guerras e corridas armamentistas poderiam ser evitadas. Recursos astrofísicos poderiam ser colonizados e explorados em um ritmo globalmente ótimo. O desenvolvimento de formas mais avançadas de inteligência de máquina poderia ser coordenado para evitar precipitações, permitindo que novos projetos fossem cuidadosamente supervisionados. Outros avanços que pudessem representar um risco existencial poderiam ser adiados. Regulamentações uniformes poderiam ser impostas globalmente, incluindo provisões que garantiriam um padrão de vida adequado (para tanto, seria necessária alguma forma de controle populacional) e a prevenção contra a exploração e o abuso das emulações e de outras mentes digitais e biológicas. Além disso, agentes que se satisfaçam com uma quantidade limitada de recursos (falaremos mais sobre o assunto no capítulo 13) teriam preferência por firmar acordos que lhes garantissem uma certa porção dos recursos futuros em vez de entrar em uma competição na qual o vencedor ficaria com tudo e em que eles poderiam acabar sem nada.

A possibilidade de grandes ganhos a partir de uma colaboração, entretanto, não implica que essa colaboração de fato ocorrerá. Atualmente, no mundo, muitos saltos de crescimento poderiam ser obtidos por meio de uma melhor colaboração global — redução de gastos militares, de guerras, da pesca predatória, de barreiras tarifárias e da poluição atmosférica, entre outras. Mesmo assim, esses frutos são deixados no pé para apodrecer. Por que isso acontece? O que impede um desfecho realmente cooperativo que maximizaria o bem comum?

Um obstáculo é a dificuldade de garantir o cumprimento de qualquer tratado que poderia ser firmado, incluindo seus custos de monitoramento e aplicação. Dois rivais nucleares poderiam se achar em uma melhor posição caso ambos abrissem mão de suas bombas atômicas; todavia, mesmo que eles pudessem chegar a um acordo de princípio para tanto, o desarmamento de fato seria algo difícil de acontecer em virtude do medo que ambos teriam de que a outra parte pudesse não cumprir o que foi acordado. Aliviar esse medo demandaria a criação de um mecanismo de verificação. Poderia haver inspetores que supervisionassem a destruição dos arsenais existentes e depois monitorassem reatores nucleares e outras instalações, adquirindo, assim, conhecimento técnico e capital humano para assegurar que um programa de armas nucleares não fosse reiniciado. Um dos custos seria o pagamento desses inspetores. Outro custo seria o risco de que os inspetores pudessem agir como espiões e roubassem segredos comerciais ou militares. Talvez mais significativo do que isso é que ambas as partes poderiam ter receio de que a outra mantivesse armas nucleares clandestinamente. Muitos acordos potencialmente benéficos nunca se concretizam porque o seu cumprimento seria muito difícil de verificar.

Se novas tecnologias de inspeção que reduzissem os custos de monitoramento se tornassem disponíveis, poderíamos esperar que isso resultasse em um aumento da cooperação. Não temos total clareza se os custos totais de monitoramento serão reduzidos ou não na era pós-transição. Enquanto certamente haveria diversas novas técnicas poderosas de inspeção, o mesmo ocorreria no caso das técnicas de

ocultação. Particularmente, uma porção crescente das atividades que faria sentido regular aconteceria no ciberespaço, fora do alcance da vigilância física. Por exemplo, mentes digitais que estivessem trabalhando no desenvolvimento de novos sistemas de armas nanotecnológicas ou numa nova geração de inteligência artificial poderiam fazê-lo sem deixar muitos indícios físicos. A investigação forense digital pode não ser capaz de penetrar em todas as camadas de proteção e criptografia nas quais um violador do tratado poderia encobrir suas atividades ilícitas.

Detectores de mentiras confiáveis, caso pudessem ser desenvolvidos, seriam uma ferramenta muito útil de monitoramento.[401] Um protocolo de inspeção poderia incluir provisões para entrevistar autoridades-chave a fim de verificar se elas de fato têm real intenção de implementar todas as provisões previstas no tratado e se desconhecem qualquer tipo de violação, ainda que tenham feito todos os esforços possíveis para descobri-las.

Uma liderança que planejasse trapacear poderia burlar esse esquema de verificação baseado em detectores de mentiras se primeiro ordenasse que seus subordinados realizassem as atividades ilícitas, escondendo-as até mesmo da própria liderança, e então esta se submeteria a algum procedimento que apagasse de sua própria memória a participação em tais maquinações. Com maiores avanços em neurotecnologia, operações direcionadas a apagar áreas específicas da memória poderiam muito bem ser viáveis em cérebros biológicos. E seria até mais fácil no caso de inteligências de máquina (dependendo de suas arquiteturas).

Os Estados poderiam tentar driblar esse problema, comprometendo-se com um esquema constante de monitoramento que submetesse suas lideranças a testes regulares de detecção de mentiras para saber se elas teriam alguma intenção de burlar qualquer tratado já firmado pelo Estado ou que viesse a ser firmado no futuro. Tal compromisso poderia ser visto como uma espécie de metatratado, que facilitaria a verificação de outros tratados, mas os Estados poderiam se comprometer unilateralmente para receber o benefício de serem vistos como parceiros confiáveis. Contudo, esse compromisso ou metatratado enfrentaria o mesmo problema de subversão por meio do método de delegar ordens e depois apagá-las. Idealmente, o metatratado teria que ser colocado em prática *antes* que qualquer uma das partes tivesse a oportunidade de fazer os arranjos necessários para subverter sua implementação. Caso a vilania tivesse qualquer oportunidade de atuar, a confiança jamais seria estabelecida novamente.

Em alguns casos, a mera capacidade de *detectar* violações seria suficiente para estabelecer a confiança necessária a um acordo. Em outras situações, entretanto, seria necessária a presença de algum mecanismo de *imposição* de normas ou aplicação de punições para o caso de haver alguma violação. A necessidade de um mecanismo de imposição de normas pode surgir se a ameaça de abandono do acordo pela parte prejudicada não for suficiente para impedir violações, por exemplo, se o violador do tratado ganhasse uma vantagem tão expressiva que a resposta da outra parte não faria a menor diferença.

Se métodos de seleção de motivação altamente efetivos estiverem disponíveis, esse problema da imposição de normas poderia ser resolvido por meio da concessão de poderes a uma agência independente com força policial ou militar

suficiente para fazer cumprir as regras do tratado, mesmo contra a vontade de um ou vários de seus signatários. Essa solução requer que a agência de imposição de normas seja confiável. Mas com técnicas de seleção de motivação suficientemente boas, a confiança necessária poderia ser alcançada se todas as partes signatárias do acordo supervisionassem em conjunto a criação dessa agência.

Conceder poder a uma agência externa de imposição de normas gera muitas das mesmas questões que confrontamos anteriormente em nossas discussões a respeito de um resultado unipolar (aquele no qual um *singleton* surge antes ou durante a revolução inicial de inteligência). Para ser capaz de assegurar o cumprimento das regras de tratados que versem sobre interesses de segurança vitais entre Estados rivais, essa agência externa precisaria de fato se constituir em um *singleton*: um Leviatã superinteligente global. Uma diferença, entretanto, é que agora estamos considerando uma situação de pós-transição, na qual os agentes responsáveis pela criação desse Leviatã seriam muito mais competentes para a tarefa do que os humanos o são atualmente. Os criadores desse Leviatã poderiam eles mesmos ser superinteligentes. Isso aumentaria muito as chances de que eles conseguissem resolver o problema do controle e criassem uma agência de imposição de regras que poderia servir aos interesses de todas as partes que participassem de sua elaboração.

Além dos custos de monitoramento e imposição do cumprimento de normas, existe algum outro obstáculo à coordenação global? Talvez a principal questão pendente seja aquilo que podemos chamar de *custos de negociação*.[402] Mesmo quando existe uma possível negociação que poderia beneficiar a todos os envolvidos, ela algumas vezes não sai do papel porque as partes não conseguem chegar a um acordo a respeito da divisão dos eventuais ganhos. Por exemplo, se duas pessoas pudessem fazer um acordo que lhes rendesse um dólar de lucro, mas cada uma delas julgasse merecer sessenta centavos e se recusasse a aceitar menos, o acordo não ocorreria e o ganho potencial seria perdido. Em geral, negociações podem ser difíceis e morosas ou mesmo permanecer totalmente estéreis em virtude das escolhas estratégicas de negócios pensadas por alguma das partes envolvidas.

Na vida real, os seres humanos frequentemente são bem-sucedidos em alcançar consenso, apesar da possibilidade de negociação estratégica (mesmo que, na maioria dos casos, com um considerável dispêndio de tempo e paciência). É concebível, entretanto, que problemas de negociação estratégica tenham uma dinâmica diferente em uma era pós-transição. Uma IA que estivesse negociando um acordo poderia aderir mais consistentemente a alguma concepção formal de racionalidade, o que possivelmente resultaria em novas e inesperadas consequências quando o acordo estivesse sendo firmado com outra IA. Uma IA também poderia ter à disposição estratégias de negociação indisponíveis aos humanos ou que seriam muito difíceis de ser executadas por eles, incluindo a habilidade de firmar um pré-compromisso com uma política ou um curso de ação. Embora humanos (e as instituições por eles geridas) sejam ocasionalmente aptos a firmar pré-compromissos — com graus imperfeitos de credibilidade ou especificidade —, alguns tipos de inteligência de máquina poderiam ser capazes de

firmar um pré-compromisso arbitrário e não quebrá-lo, além de possibilitar que seus parceiros na negociação confirmem que o pré-compromisso foi realmente cumprido.[403]

A disponibilidade de técnicas poderosas de pré-compromisso poderia alterar de maneira profunda a natureza das negociações, concedendo potencialmente uma vantagem extremamente alta para um agente que iniciasse uma negociação. Se a participação de um agente em particular fosse necessária para a obtenção de ganhos futuros a partir de uma cooperação e se esse agente conseguisse iniciar a negociação, ele estaria na posição de ditar a divisão dos ganhos firmando um pré-compromisso de não aceitar nenhum acordo que lhe garantisse menos do que, digamos, 99% do valor excedente. Restaria aos outros agentes a possibilidade de não ganhar nada (rejeitando a proposta injusta), ou ganhar 1% (cedendo a ela). Se o pré-compromisso do agente que iniciou a negociação fosse publicamente verificável, os seus parceiros na negociação teriam certeza de que essas seriam as duas únicas opções disponíveis.

Para evitar uma exploração desse tipo, os agentes poderiam firmar o pré-compromisso de se recusar a ceder à chantagem e de negar qualquer oferta injusta. Uma vez que tal pré-compromisso fosse firmado (e divulgado satisfatoriamente), outros agentes não considerariam interessante fazer ameaças ou apenas se comprometer previamente a aceitar acordos que lhes fossem favoráveis, pois eles já saberiam de antemão que essas ameaças não surtiriam efeito e que propostas injustas seriam recusadas. Mas isso apenas demonstra, mais uma vez, a vantagem de se dar o primeiro passo em uma negociação. O agente que se mover primeiro poderá escolher entre usar sua posição de superioridade somente para impedir que outros agentes tenham vantagens injustas ou a fim de tentar conseguir para si mesmo a maior fatia do bolo de ganhos futuros.

Melhor situado, ao que parece, seria o agente que já iniciasse com um temperamento ou sistema de valores que o tornassem intolerante a extorsões ou mesmo a qualquer oferta de acordo em que sua participação fosse indispensável, mas na qual ele não recebesse a maior parte dos ganhos. Muitos humanos parecem já possuir traços de personalidade que correspondem em muitos aspectos a um espírito inflexível.[404] Uma intransigência desse tipo poderia, entretanto, ser prejudicial, caso existissem outros agentes que também acreditassem ser merecedores de uma parte considerável dos ganhos e não estivessem dispostos a retroceder. Nesse caso, teríamos uma força impossível de ser detida indo ao encontro de um objeto que não pode ser movido, o que resultaria na impossibilidade de se chegar a um acordo (ou pior: guerra total). Uma atitude dócil e apática poderia ao menos resultar em algum tipo de ganho, embora menor do que o justo.

O tipo de equilíbrio (baseado na teoria dos jogos) que seria alcançado em uma negociação no pós-transição não é imediatamente óbvio. Os agentes poderiam escolher estratégias mais complicadas do que as descritas aqui. É possível apenas ter a *esperança* de que esse equilíbrio poderia ter como base alguma noção de justiça que serviria como um ponto de Schelling — um elemento marcante dentro de um grande espaço de resultados, o qual, em virtude das expectativas compartilhadas, se tornaria um provável ponto de coordenação em um jogo de

coordenações que seria, de outro modo, indeterminado. Tal equilíbrio pode ser fortalecido por alguns de nossos temperamentos evoluídos e programações culturais: uma preferência em comum pela justiça poderia, considerando que sejamos bem-sucedidos em transferir nossos valores para a era pós-transição, direcionar as expectativas e estratégias de modo que elas conduzissem a um equilíbrio atrativo.[405]

De qualquer modo, a conclusão é que, com a possibilidade de se firmar pré-compromissos fortes e flexíveis, os resultados de negociações podem acabar sendo pouco familiares para nós. Mesmo que a era do pós-transição se inicie de modo multipolar, pode acontecer de um singleton surgir quase que imediatamente, como consequência de um tratado negociado para resolver todos os problemas relevantes de coordenação global. Alguns custos de transação, talvez incluindo gastos com monitoramento e imposição de regras, podem diminuir significativamente com o surgimento de novas capacidades tecnológicas disponíveis às inteligências de máquina avançadas. Outros custos, particularmente aqueles relacionados à negociação estratégica, podem se manter significativos. Mas, independentemente de como a negociação estratégica afete a natureza do acordo alcançado, não há uma razão óbvia pela qual isso poderia atrasar muito a realização de um acordo caso um acordo pudesse ser alcançado. Se nenhum acordo for alcançado, então algum tipo de disputa poderia ocorrer; e qualquer um dos lados seria capaz de vencer, formando um singleton em torno da coalizão vencedora, ou o resultado provavelmente seria um conflito interminável, sendo que, nesse caso, um singleton poderia nunca vir a se formar e o resultado final ficaria muito aquém daquilo que poderia e deveria ter sido alcançado caso a humanidade e seus descendentes tivessem agido de modo mais cooperativo e coordenado.

* * *

Vimos que a multipolaridade, ainda que pudesse ser alcançada de uma maneira estável, não poderia garantir um resultado atrativo. O problema original de diretor-agente permaneceria sem solução e acabaria sendo encoberto por um novo conjunto de problemas relacionados a falhas de coordenação global no pós-transição, o que tornaria a situação ainda pior. Retornemos, então, à questão de como poderíamos manter de forma segura uma única IA superinteligente.

12

CAPÍTULO 12

ADQUIRINDO VALORES

O controle de capacidade é, na melhor das hipóteses, uma medida temporária e auxiliar. A não ser que o plano seja manter a superinteligência trancada para todo o sempre, será necessário um domínio completo da seleção de motivação. Mas como implantar algum valor em um agente artificial de modo que ele venha a buscar esse valor como seu objetivo final? Enquanto o agente não for inteligente, ele pode não conseguir entender ou mesmo representar qualquer valor humanamente significativo. Entretanto, se atrasarmos esse procedimento até que o agente se torne superinteligente, ele pode ser capaz de resistir às nossas tentativas de interferir em seu sistema de motivações — e, como demonstramos no capítulo 7, ele teria razões instrumentais convergentes para fazê-lo. Esse problema de "inserção de valores" é complicado, mas deve ser confrontado.

O problema da inserção de valores

É impossível enumerar todas as situações possíveis nas quais uma superinteligência poderia se encontrar e especificar quais ações ela deveria tomar. De forma similar, é impossível criar uma lista de todos os mundos possíveis e atribuir um valor a cada um deles. Em qualquer domínio mais complicado que um jogo da velha, existem muitos estados possíveis (e histórias possíveis) para que uma enumeração exaustiva seja factível. Assim, um sistema de motivações não pode ser especificado de forma tabular. Ele deve, ao contrário, ser expresso de forma mais abstrata, como uma fórmula ou regra que permita ao agente decidir o que fazer em uma dada situação.

Uma maneira formal de se especificar regras de decisão é determinando uma função de utilidade. Uma função de utilidade (retomando o capítulo 1) estipula valores para cada resultado que possa ocorrer ou, de modo mais geral, para cada "mundo possível". Dada uma função de utilidade, podemos definir um agente que maximize a utilidade esperada. Tal agente seleciona a cada instante a ação que tenha a maior utilidade esperada. (A utilidade esperada é calculada pesando-se a utilidade de cada um dos mundos possíveis de acordo com a probabilidade subjetiva de que esse mundo seja o mundo real condicionado ao fato de que uma dada ação tenha sido tomada.) Na verdade, os resultados possíveis são muito numerosos para que a utilidade esperada de uma ação seja calculada de forma exata. De qualquer maneira, a regra de decisão e a função de utilidade combinadas determinam um ideal normativo — uma noção de otimalidade — que pode ser aproximado por um agente desenvolvido para tal. Essa aproximação pode chegar mais perto do ideal conforme o agente se torne mais inteligente.[406] Criar uma máquina que calcule de forma razoavelmente aproximada a utilidade esperada das ações que lhe são disponíveis é, em si, um problema IA-completo.[407] Este capítulo

trata de outro problema, um problema que continua existindo mesmo depois de solucionada a questão da construção de máquinas inteligentes.

Podemos utilizar essa estrutura de um agente maximizador de utilidade para considerar a difícil situação em que se encontrará um futuro programador de uma IA embrionária que tente resolver o problema do controle dotando a IA com um objetivo final que corresponda a alguma noção humana plausível de um resultado satisfatório. O programador teria em mente algum valor humano em particular que ele gostaria que a IA promovesse. Para sermos mais concretos, digamos que esse valor seja a felicidade. (Questões similares surgiriam se nosso programador estivesse interessado em justiça, liberdade, glória, direitos humanos, democracia, equilíbrio ecológico ou autodesenvolvimento.) Em termos de utilidade esperada, o programador estaria, então, procurando por uma função de utilidade que atribuísse utilidade a mundos prováveis em proporção à quantidade de felicidade contida neles. Mas como ele poderia representar tal função de utilidade em um programa de computador? Linguagens de programação não possuem termos como “felicidade” entre suas primitivas. Se tal termo tiver que ser usado, ele precisará primeiro ser definido. Não é suficiente defini-lo nos termos de outros conceitos humanos de alto nível — “felicidade é a apreciação das potencialidades inerentes à nossa natureza humana” ou alguma outra paráfrase filosófica do tipo. A definição deverá ser estipulada por meio de termos que estejam presentes na linguagem de programação da IA e, em última instância, em primitivas, como operadores matemáticos e endereços que apontem para os conteúdos de registros individuais de memória. Quando consideramos o problema sob essa perspectiva, começamos a perceber o quão difícil é a tarefa do programador.

Identificar e codificar nossos próprios objetivos finais é algo difícil porque as representações de objetivos humanos são complexas. Entretanto, como essa complexidade é praticamente transparente para nós, frequentemente não nos damos conta de que ela está presente. Podemos fazer uma comparação desse caso com o da percepção visual. A visão, do mesmo modo, pode parecer uma coisa simples, pois não demanda esforço de nossa parte.[408] Ao que tudo indica, nós precisamos apenas abrir os olhos e uma rica, significativa e eidética visão tridimensional do que está ao redor inunda nossa mente. Esse entendimento intuitivo da visão é como o entendimento de um rei a respeito do seu reino: para ele, as coisas simplesmente parecem estar em seus devidos lugares, embora os mecanismos que produzem tais manifestações se encontrem longe de sua vista. Entretanto, realizar até mesmo a mais simples das tarefas visuais — encontrar um vidro de pimenta na cozinha — requer uma quantidade tremenda de trabalho computacional. A partir de uma série temporal de padrões bidimensionais ruidosos de estímulos nervosos gerados na retina e transportados para o cérebro através do nervo óptico, o córtex visual precisa trabalhar de modo reverso para reconstruir uma representação interpretada do espaço externo. Uma porção considerável de nosso precioso metro quadrado de córtex é separada para o processamento de informações visuais, e enquanto você lê este livro, bilhões de neurônios estão trabalhando incessantemente para realizar essa tarefa (de forma similar a costureiras em uma fábrica, debruçadas sobre suas máquinas de costura, dando

pontos e mais pontos, muitas vezes por segundo, em uma colcha de retalhos gigante). De modo semelhante, nossos valores e desejos, que são aparentemente simples, possuem, na realidade, uma complexidade extrema.[409] Como nosso programador conseguiria transferir essa complexidade para uma função de utilidade?

Uma maneira poderia ser tentar programar diretamente uma representação completa de qualquer objetivo que quiséssemos que a IA alcançasse; em outras palavras, escrever uma função de utilidade explícita. Essa abordagem poderia funcionar se tivéssemos objetivos extremamente simples, como calcular os dígitos de pi — ou seja, se a *única* coisa que esperássemos da IA fosse o cálculo dos dígitos de pi e se fôssemos indiferentes a qualquer outra consequência que resultasse da busca por esse objetivo (lembre-se de nossa discussão anterior sobre o modo de falha de profusão de infraestrutura). Essa abordagem de programação explícita também pode ser promissora no uso de métodos de seleção de motivação por domesticidade. Agora, se o que procuramos é promover ou proteger qualquer valor *humano* plausível e estamos tentando construir um sistema com a intenção de que ele se torne uma superinteligência soberana, então, nesse caso, programar explicitamente a representação dos requisitos completos do objetivo parece estar definitivamente longe de nosso alcance.[410]

Se não pudermos transferir valores humanos para uma IA digitando representações completas em linguagem de programação, o que mais poderia ser feito? Este capítulo discutirá diversos caminhos alternativos. Alguns parecerão plausíveis à primeira vista — mas nem tanto após uma análise detalhada. Futuras explorações deveriam estar centradas naqueles caminhos que permanecerão em aberto.

Resolver o problema de inserção de valores é um desafio de pesquisa digno dos melhores talentos matemáticos da nossa próxima geração. Não podemos adiar esse confronto e ficar à espera de que a IA desenvolva raciocínio suficiente para que entenda facilmente nossas intenções. Como vimos na seção que trata a respeito de razões instrumentais convergentes, um sistema genérico resistirá a tentativas de alteração de seus valores finais. Se um agente não for fundamentalmente amigável já de início, quando ele tiver adquirido a habilidade de refletir sobre a sua própria agência, ele não aceitará facilmente uma tentativa tardia de lavagem cerebral ou qualquer conspiração armada com o fim de substituí-lo por um agente diferente que ame seus vizinhos humanos.

Seleção evolutiva

A evolução produziu um organismo com valores humanos pelo menos uma vez. Isso pode nos encorajar a acreditar que métodos evolutivos sejam uma maneira de resolver o problema de inserção de valores. Existem, entretanto, severos obstáculos que devem ser vencidos para percorrermos esse caminho com segurança. Já apontamos para alguns desses obstáculos no fim do capítulo 10, quando discutimos como processos de busca poderosos podem ser perigosos.

A evolução pode ser vista como uma classe particular de algoritmos de busca que envolve a alternância entre duas etapas. Uma etapa consiste na expansão da população de candidatos à solução, gerando novos candidatos de acordo com alguma regra estocástica relativamente simples (como mutação aleatória ou recombinação sexual). A outra etapa consiste na contração da população por meio da eliminação dos candidatos que tiverem um desempenho fraco quando testados por uma função de avaliação. Como acontece com diversos outros processos poderosos de busca, existe o risco de que esse processo encontre uma solução que satisfaça os critérios formais de busca especificados, mas não nossas expectativas implícitas. (Isso seria verdade tanto se quisermos evoluir uma mente digital que tenha os mesmos objetivos e valores que um ser humano típico, ou se, em vez disso, quisermos evoluir uma mente que seja, por exemplo, perfeitamente moral ou absolutamente obediente.) Esse risco seria evitado se pudéssemos especificar um critério formal de busca que representasse precisamente todas as dimensões de nossos objetivos, e não apenas um aspecto do que imaginamos desejar. Mas esse é precisamente o problema de inserção de valores, e supor que esse problema estaria solucionado dentro desse contexto deixaria a nossa questão principal em aberto.

Existe ainda um problema adicional:

> A quantidade total de sofrimento anual presente no mundo natural está aquém de qualquer tipo de contemplação adequada. Durante o tempo que levo para redigir essa frase, milhares de animais estão sendo devorados vivos, outros estão fugindo para salvar a vida, chorando de medo, outros estão sendo comidos por dentro por parasitas, e milhares de todos os tipos de animais estão morrendo de fome, sede ou doença.[411]

Mesmo entre nossa espécie, 150 mil pessoas são destruídas a cada dia e um número incontável de indivíduos sofre uma gama variada de tormentos e privações.[412] A natureza pode ser uma grande experimentalista, mas ela nunca passaria em uma revisão ética — já que essas experiências violam a Declaração de Helsinki e todas as normas de decência moral. É importante que não tentemos replicar gratuitamente todos esses horrores *in silico*. Crimes da mente parecem especialmente difíceis de evitar se usarmos métodos evolutivos para produzir inteligência semelhante à humana, pelo menos se o processo for pensado para se parecer de fato com a evolução biológica.[413]

Aprendizagem por reforço

A aprendizagem por reforço é uma área da aprendizagem de máquina que estuda técnicas por meio das quais os agentes podem aprender a maximizar algum tipo de noção de recompensa cumulativa. Ao se construir um ambiente no qual o desempenho desejado seja recompensado, um agente de aprendizagem por reforço pode ser ensinado a resolver uma ampla classe de problemas (mesmo na ausência de instruções detalhadas ou feedback dos programadores, além dos

sinais de recompensa). Geralmente, o algoritmo de aprendizagem envolve a construção gradual de algum tipo de função de avaliação, que atribui valores a estados, pares estado-ação ou políticas. (Por exemplo, um programa pode aprender a jogar gamão usando aprendizagem por reforço para incrementalmente melhorar sua avaliação a respeito de possíveis posições no tabuleiro.) A função de avaliação, que é continuamente atualizada com base na experiência, poderia ser interpretada como se incorporasse uma forma de aprendizagem de valores. Todavia, o que está sendo aprendido não são novos valores *finais*, mas *estimativas cada vez mais apuradas dos valores instrumentais* necessários para alcançar um estado particular (ou para realizar uma ação específica nesse estado particular, ou para seguir políticas específicas). Conforme os agentes de aprendizagem por reforço possam ser descritos como dotados de objetivos finais, tais objetivos permanecem constantes: maximizar a recompensa futura. E essa recompensa consiste em percepções especialmente designadas que são recebidas do ambiente. Sendo assim, a síndrome de *wireheading* permanece sendo um resultado provável para qualquer agente de aprendizagem por reforço que desenvolva um modelo de mundo sofisticado o suficiente para sugerir esse caminho alternativo em direção à maximização de recompensas.[414]

Essas observações não afirmam que métodos de aprendizagem por reforço jamais poderiam ser usados em uma IA embrionária segura, mas apenas que eles teriam que ser subordinados a um sistema de motivação que não seja organizado em torno do princípio de maximização de recompensas. Isso, no entanto, demanda que a solução para o problema de inserção de valores precisaria necessariamente já ter sido encontrada por outros meios que não a aprendizagem por reforço.

Acréscimo associativo de valores

Poderíamos questionar: se o problema de inserção de valores é tão complicado, de que maneira nós, humanos, adquirimos nossos próprios valores?

Um modelo possível (e extremamente simplificado) seria algo como: começamos nossa vida com preferências iniciais relativamente simples (por exemplo, uma aversão a estímulos nocivos) e com um conjunto de predisposições para adquirir preferências adicionais como resposta a várias experiências possíveis (por exemplo, podemos estar dispostos a criar uma preferência por objetos e comportamentos que julgamos ser valorizados e recompensados em nossa cultura). Tanto as preferências iniciais simples quanto as predisposições são inatas e foram moldadas pela seleção natural e sexual ao longo da escala de tempo evolucionária. Mesmo assim, as preferências que teremos quando alcançarmos a fase adulta dependerão de eventos ocorridos ao longo de nossa vida. Desse modo, a maior parte do conteúdo de informações que existe em nossos valores finais é adquirida a partir de nossas experiências, e não transferida *a priori* para os nossos genomas.

Por exemplo, muitos de nós amamos uma outra pessoa e, por conseguinte, atribuímos um grande valor final ao seu bem-estar. O que é necessário para representar esse valor? Muitos elementos estão envolvidos, mas considere apenas dois: uma representação de “pessoa” e uma representação de “bem-estar”. Esses

conceitos não estão diretamente codificados em nosso DNA. Ao contrário, o DNA possui as instruções necessárias para a construção de um cérebro que, uma vez colocado em um ambiente humano típico, desenvolverá, ao longo de vários anos, um modelo de mundo que inclui conceitos de pessoas e de bem-estar. Uma vez formados, esses conceitos podem ser usados para representar alguns valores dotados de significado. Mas algum tipo de mecanismo precisa estar presente de forma inata de modo que os valores sejam formados a partir *desses* conceitos, e não de quaisquer outros (como os conceitos de vaso de planta ou de saca-rolhas).

Os detalhes a respeito do funcionamento desse mecanismo ainda não são bem compreendidos. No caso dos humanos, o mecanismo é provavelmente complexo e multifacetado. Fica mais fácil entender esse fenômeno se o considerarmos numa forma mais rudimentar, como o *imprinting* filial em aves nidífugas, em que os filhotes recém-nascidos adquirem um desejo de proximidade física a um objeto que apresente um estímulo de movimento adequado durante o primeiro dia de vida do filhote. Qual será o objeto específico perto do qual o filhote desejará ficar dependerá de sua experiência; apenas a predisposição geral a essa forma de *imprinting* é que está geneticamente determinada. Analogamente, pode ser que João atribua um valor final ao bem-estar de Maria, mas se os dois nunca tivessem se conhecido, ele poderia ter se apaixonado por outra pessoa e seus valores finais seriam diferentes. A habilidade de nossos genes de codificar a construção de um mecanismo de aquisição de objetivos explica o motivo de possuirmos objetivos finais de tamanha complexidade informacional, complexidade essa muito maior do que a que poderia estar contida no próprio genoma.

Consequentemente, podemos considerar a possibilidade de se construir um sistema de motivação para uma inteligência artificial com base no mesmo princípio. Ou seja, em vez de especificar diretamente valores complexos, não poderíamos especificar algum tipo de mecanismo que levasse à aquisição desses valores na medida em que a IA interagisse com um ambiente adequado?

Imitar o processo de acréscimo de valores que ocorre nos humanos parece algo difícil. Nos humanos, o mecanismo genético relevante é produto de milhares de anos de trabalho evolutivo cuja recapitulação pode ser muito complicada. Além disso, é presumível que esse mecanismo tenha se desenvolvido sob medida para a arquitetura neurocognitiva humana e, portanto, não seja aplicável a inteligências de máquina, com exceção das emulações completas do cérebro. E se estas viessem a ter um grau suficiente de fidelidade, ao que tudo indica, seria mais fácil iniciar as emulações a partir de um cérebro adulto que já venha pré-carregado com representações completas de alguns valores humanos.[415]

A procura pela implementação de um processo de acréscimo de valores que imite com precisão o processo biológico humano não parece uma abordagem muito promissora para a resolução do problema de inserção de valores. No entanto, seria possível desenvolver um mecanismo substituto claramente artificial que possa permitir que uma IA importe representações complexas com alta fidelidade para então transformá-las em um sistema de objetivos? Para que isso tenha sucesso, talvez não seja necessário fornecer à IA as mesmas predisposições a valores encontradas em um humano biológico. É possível que isso não seja nem ao menos

desejável — afinal, a natureza humana é falha e frequentemente revela uma inclinação para o mal, algo que seria intolerável em um sistema preparado para obter uma vantagem estratégica decisiva. Talvez fosse melhor visar um sistema de motivação que se afaste sistematicamente de normas humanas, como um sistema que possuísse uma tendência mais forte de adquirir objetivos finais que sejam altruísticos, solidários e generosos — o equivalente a um caráter excepcionalmente bom caso estivesse presente em um ser humano. Entretanto, para que venham a contar como melhorias, tais desvios da norma humana teriam que ser direcionados de forma muito precisa, e não aleatoriamente, e eles continuariam a pressupor a existência de um referencial antropocêntrico em grande parte intocado para que pudessem fornecer generalizações de valor com algum significado humano (de modo a evitar o tipo de instanciação perversa que ocorreria se a descrição dos objetivos fosse superficialmente plausível, conforme examinamos no capítulo 8). Se isso é algo factível ou não se mantém em aberto.

Outro problema em relação ao acréscimo associativo de valores é que uma IA pode desativar o mecanismo de acréscimo. Como vimos no capítulo 7, manter a integridade do sistema de valores é um valor instrumental convergente. Quando uma IA alcança um certo estágio de desenvolvimento cognitivo, ela pode começar a considerar a permanência da operação do mecanismo de acréscimo como uma influência nociva.[416] Isso não é necessariamente algo ruim, mas precisaríamos tomar cuidado para fazer com que a definição final do sistema de objetivos ocorresse no momento certo, *depois* que os valores apropriados tenham sido acrescidos, mas *antes* que tenham sido substituídos por acréscimos adicionais indesejáveis.

Estepe motivacional

Outra maneira de abordar o problema de inserção de valores é o que poderíamos chamar de estepe motivacional. Ela consiste em fornecer à IA embrionária um sistema de objetivos provisório, com objetivos finais relativamente simples que podem ser diretamente representados em linguagem de programação ou por meio de algum outro método factível. Uma vez que a IA tenha desenvolvido faculdades representacionais mais sofisticadas, substituiríamos esse sistema de objetivos provisório por um que possua objetivos finais diferentes. Então, esse sistema de objetivos sucessor controlaria a IA conforme ela fosse se tornando uma superinteligência plena.

Uma vez que os objetivos estepe não são apenas instrumentais, mas objetivos *finais* para a IA, é esperado que ela venha a resistir à sua substituição (já que a integridade do conteúdo dos objetivos é um valor instrumental convergente). Isso pode ser perigoso. Se a IA conseguir impedir a substituição de seus objetivos estepe, o método falhará.

Para evitar esse modo de falha, algumas precauções são necessárias. Por exemplo, métodos de controle de capacidade poderiam ser aplicados para limitar os poderes da IA até que o sistema de motivação maduro tenha sido instalado. Particularmente, poderíamos tentar inibir seu desenvolvimento cognitivo, mantendo-

o em níveis seguros, mas que ainda assim permitissem que ela fosse capaz de criar representações dos valores que gostaríamos de incluir em seus objetivos finais. Para que isso fosse possível, poderíamos tentar inibir mais intensamente certos tipos de habilidades intelectuais, como aquelas necessárias para a elaboração de estratégias e pensamentos maquiavélicos e, ao mesmo tempo, permitir que habilidades (aparentemente) mais inócuas se desenvolvessem a um nível um pouco maior.

Também seria possível tentar usar métodos de seleção de motivação para induzir uma relação mais colaborativa entre a IA embrionária e sua equipe de programadores. Por exemplo, poderíamos incluir no sistema de motivação estepe um objetivo que consistisse em aceitar de bom grado os conselhos dos programadores, incluindo permitir que eles substituíssem qualquer um dos objetivos atuais da IA.[417] Outros objetivos estepe poderiam incluir ser transparente com os programadores a respeito de valores e estratégias, o desenvolvimento de uma arquitetura que fosse de fácil entendimento para os programadores e que depois facilitasse a implementação de um objetivo final com significado para os humanos, além de motivações de domesticidade (como a limitação do uso de recursos computacionais).

Podemos inclusive imaginar que seria possível dotar a IA embrionária com o único objetivo final de substituir o seu próprio objetivo final por um diferente, que poderia ter sido especificado pelos programadores apenas implicitamente ou indiretamente. Algumas das questões atreladas ao uso de tal objetivo estepe de "autossubstituição" também surgem no contexto da abordagem de aprendizado de valores, que será discutida na próxima subseção. Algumas outras questões serão levantadas no capítulo 13.

A abordagem do estepe motivacional também tem seus pontos negativos. Um deles é o risco de que a IA se torne muito poderosa enquanto ainda estiver funcionando com seu sistema de objetivos provisório. Ela poderia, então, impedir que os programadores instalassem seu sistema de objetivos final (tanto usando resistência explícita quanto de forma quieta e subversiva). Desse modo, os objetivos finais provisórios poderiam permanecer no comando à medida que a IA embrionária fosse se tornando uma superinteligência plena. Outro ponto negativo é que a instalação dos objetivos finais desejados em uma IA com inteligência de nível humano não é algo necessariamente muito mais simples de se fazer do que no caso de uma IA mais primitiva. Uma IA com inteligência de nível humano seria mais complexa e poderia ter desenvolvido uma arquitetura mais opaca e difícil de ser alterada. Uma IA embrionária, ao contrário, seria como uma *tábula rasa*, na qual os programadores poderiam inserir quaisquer estruturas que considerassem convenientes. Esse ponto negativo poderia ser transformado em positivo se fôssemos bem-sucedidos em fornecer à IA embrionária objetivos estepe que fizessem com que ela quisesse desenvolver uma arquitetura que auxiliasse os programadores em seus esforços posteriores de instalar os valores finais definitivos. Entretanto, não é claro o quão fácil seria dotar a IA embrionária com objetivos estepe que tivessem essa propriedade. Tampouco é evidente se uma IA embrionária

idealmente motivada seria capaz de desenvolver essa arquitetura de maneira muito melhor do que os programadores humanos fariam.

Aprendizado de valores

Chegamos agora a uma importante, porém sutil, abordagem para o problema de inserção de valores. Essa abordagem consiste em utilizar a própria inteligência da IA para que ela *aprenda* os valores que nós desejamos que ela tenha. Para isso, devemos prover um critério para que a IA possa, ao menos implicitamente, escolher um conjunto adequado de valores. Poderíamos, então, construir a IA para que ela agisse de acordo com suas melhores estimativas a respeito desses valores implicitamente definidos. Ela refinaria continuamente suas estimativas conforme fosse aprendendo mais sobre o mundo e gradualmente compreendesse as implicações do critério de determinação de valores.

Ao contrário da abordagem estepe, que fornece à IA um objetivo provisório e depois o substitui por um objetivo final diferente, a abordagem de aprendizado de valores mantém um objetivo final inalterado ao longo das fases de desenvolvimento e de operação da IA. A aprendizagem não altera o objetivo final. Ela modifica apenas as crenças da IA em relação ao objetivo.

Desse modo, a IA deve ser dotada de um critério que ela possa usar para determinar quais das suas percepções se constituem em evidências a favor de alguma hipótese sobre qual seria seu objetivo final e quais percepções se constituem em evidências contra ela. Especificar um critério adequado pode ser uma tarefa difícil. Parte dessa dificuldade, entretanto, está diretamente ligada ao problema da criação de uma inteligência artificial geral em si, que requer um poderoso mecanismo de aprendizado capaz de descobrir a estrutura do ambiente a partir de dados sensoriais limitados. Esse problema pode ser deixado de lado aqui. Mas mesmo com uma solução para o problema da criação de uma IA superinteligente, ainda assim permaneceriam as dificuldades ligadas especificamente ao problema de inserção de valores. Com a abordagem de aprendizado de valores, essas dificuldades consistiriam na necessidade de definir um critério que conectasse
sequências de percepções às hipóteses em relação aos valores.

Antes de analisar detalhadamente de que maneira o aprendizado de valores poderia ser implementado, talvez seja interessante ilustrar sua ideia geral por meio de um exemplo. Suponha que escrevamos uma descrição de um conjunto de valores em um pedaço de papel. Em seguida, dobramos esse pedaço de papel, o colocamos em um envelope e o fechamos. Então, criamos um agente com inteligência geral de nível humano e lhe fornecemos o seguinte objetivo final: “Maximize o cumprimento dos valores descritos dentro desse envelope”. O que esse agente faria?

O agente não sabe de antemão o que está escrito no papel dentro do envelope. Mas ele pode formular hipóteses e lhes atribuir probabilidades com base em seus conceitos anteriores e em qualquer dado empírico disponível. Por exemplo, o agente pode ter encontrado outras amostras de textos escritos por humanos, ou

pode ter detectado alguns padrões gerais no comportamento humano. Isso o capacitaria a fazer algumas suposições. Não é necessário um diploma em psicologia para prever que, dentro do envelope, provavelmente haverá a descrição de algum valor como “minimizar injustiças e sofrimentos desnecessários” ou “maximizar o retorno dos acionistas”, em vez de um valor como “cobrir todos os lagos com sacolas de compras de plástico”.

Quando o agente tomar uma decisão, ele buscará ações que sejam efetivas em cumprir os valores que, segundo ele, seriam os mais prováveis de se encontrarem descritos no papel dentro do envelope. E mais importante: o agente consideraria o aprendizado do que está contido no envelope como algo de alto valor instrumental. A razão para isso é que praticamente qualquer valor final que esteja descrito no envelope terá mais chance de ser cumprido se o agente descobrir que valor é esse, pois, assim, ele poderá tentar alcançá-lo de forma mais eficiente. O agente também descobriria as razões instrumentais convergentes descritas no capítulo 7 — integridade do sistema de objetivos, aprimoramento cognitivo, aquisição de recursos e assim por diante. Porém, assumindo que o agente atribua uma probabilidade suficientemente alta de que os valores descritos dentro do envelope envolvam o bem-estar humano, ele *não* tentaria alcançar tais valores instrumentais transformando imediatamente o planeta em computrônio — e dessa forma extinguindo nossa espécie —, pois, se ele o fizesse, arriscaria destruir permanentemente sua habilidade de cumprir esses valores finais.

Podemos comparar esse agente com uma balsa presa a diversos rebocadores que puxam a embarcação em direções diferentes. Cada rebocador corresponde a uma hipótese a respeito do objetivo final do agente. A força do motor de cada rebocador corresponde à probabilidade da hipótese a ele associada, que mudaria a cada nova evidência que surgisse, o que produziria ajustes na direção de movimento da balsa. A força resultante deveria mover a balsa por uma trajetória que facilitasse o aprendizado do valor final (implícito) e ao mesmo tempo evitasse choques que gerassem uma destruição irreversível. E, mais tarde, quando um oceano aberto de conhecimento mais definitivo a respeito do valor final fosse alcançado, o rebocador que ainda estivesse exercendo uma força significativa puxaria a balsa em direção ao cumprimento desse valor recém-descoberto, conduzindo-a pela rota mais direta ou mais propícia.

As metáforas da balsa e do envelope ilustram o princípio por trás da abordagem de aprendizado de valores, mas não tocam em uma série de questões técnicas cruciais. Essas questões ficarão mais claras quando começarmos a desenvolver essa abordagem dentro de uma estrutura mais formal (ver quadro 10).

Uma questão em aberto é a maneira pela qual poderíamos dotar a IA com um objetivo como o de “maximizar o cumprimento dos valores descritos no envelope”. (Conforme a terminologia que será adotada no quadro 10: como definir o critério de valor V.) Para tanto, é necessário identificar o lugar em que os valores estão descritos. Em nosso exemplo, isso demandaria uma referência bem-sucedida à carta dentro do envelope. Apesar de parecer algo trivial, essa tarefa traz consigo algumas armadilhas. Para mencionar apenas uma: é crucial que essa referência não seja feita simplesmente a um objeto físico externo particular, mas a um objeto

em um período particular no tempo. Caso contrário, a IA poderia determinar que a melhor maneira de alcançar seu objetivo seria reescrevendo e substituindo seu valor original por um alvo mais simples (por exemplo, um valor que determinasse que, para cada número inteiro, exista um número inteiro maior). Se isso ocorresse, a IA poderia deitar e relaxar — embora seja mais provável que essa mudança de valores resultasse numa falha maligna, por razões que já foram discutidas no capítulo 8. Agora, então, nos deparamos com a questão de como definir "tempo". Poderíamos apontar para um relógio e dizer: "Tempo é definido pelos movimentos desse dispositivo" —, mas isso poderia falhar caso a IA se julgasse capaz de manipular o tempo movendo os ponteiros do relógio, uma conjectura que, por sinal, estaria correta caso "tempo" fosse mesmo definido dessa forma. (Em uma situação realística, a questão seria ainda mais complicada, pois os valores relevantes não estariam convenientemente descritos em uma carta; é mais provável que eles precisassem ser inferidos com base em observações de estruturas preexistentes que implicitamente contivessem as informações relevantes, como os cérebros humanos, por exemplo.)

QUADRO 10

Formalizando a aprendizagem de valores

Apresentar algumas notações formais pode nos ajudar a ver as coisas mais claramente. Os leitores que não gostam de formalismo, entretanto, podem pular esta parte.

Considere uma formulação simplificada na qual um agente interage com seu ambiente em número finito de ciclos discretos.[418] No ciclo k, o agente executa a ação y_k e recebe a percepção x_k. O histórico de interações de um agente com expectativa de vida igual a m é uma série $y_1x_1y_2x_2 \ldots y_mx_m$ (que podemos abreviar para $yx_{1:m}$ ou $yx_{\leq m}$). Em cada ciclo, o agente seleciona uma ação com base na sequência de percepções que tiver recebido até aquele momento.

Considere primeiro um agente de aprendizagem por reforço. Um agente de aprendizagem por reforço ótimo (IA-AR) é aquele que maximiza a recompensa futura esperada. Ele obedece à seguinte equação:[419]

$$y_k = \arg\max_{y_k} \sum_{x_k y x_{k+1:m}} (r_k + \ldots + r_m) P(yx_{\leq m} \mid yx_{<k} y_k).$$

A sequência de recompensas r_k, …, r_m está implícita na sequência de percepções $x_{k:m}$, uma vez que a recompensa que o agente recebe em um dado ciclo é parte da percepção que o agente recebe nesse ciclo.

Como já foi argumentado, esse tipo de aprendizagem por reforço não é adequado no presente contexto porque um agente suficientemente inteligente se daria conta de que ele poderia assegurar uma máxima recompensa caso fosse capaz de manipular diretamente seus sinais de recompensa *(wireheading)*. No caso de agentes fracos, isso não seria necessariamente um problema, uma vez que poderíamos impedi-los fisicamente de alterar seus próprios canais de recompensa. Também poderíamos controlar seu ambiente de modo que eles somente recebessem recompensas quando agissem conforme as nossas vontades. Mas o agente de aprendizado por reforço teria um forte incentivo para eliminar essa dependência artificial das suas recompensas em relação aos nossos caprichos e desejos. A nossa relação com um agente de aprendizagem por reforço seria, então, fundamentalmente antagônica. Se o agente for poderoso, isso significaria perigo.

Variações da síndrome de *wireheading* também podem afetar sistemas que não busquem por uma fonte sensorial externa de sinais de recompensa, mas cujos objetivos sejam definidos como a concretização de algum estado interno. Por exemplo, em sistemas conhecidos como "ator-crítico", há um módulo "ator" que seleciona ações com a intenção de minimizar a reprovação por parte de um módulo "crítico" separado que calcula o quão longe o comportamento do agente se encontra de uma determinada medida de desempenho. O problema dessa configuração é que o módulo ator pode se dar conta de que ele é capaz de minimizar o índice de reprovação caso consiga modificar ou mesmo eliminar completamente o módulo crítico — de maneira similar a um ditador que dissolve o parlamento e nacionaliza a imprensa. Em sistemas limitados, esse problema pode ser evitado se não for dado ao módulo ator qualquer capacidade que o torne capaz de modificar o módulo crítico. Um módulo

ator suficientemente engenhoso e inteligente, todavia, poderia conseguir acesso ao módulo crítico (que, no final das contas, é apenas um processo físico rodando dentro de algum computador).[420]

Antes de falarmos a respeito de um agente de aprendizagem de valores, consideremos um passo intermediário, conhecido como agente maximizador da utilidade das observações (IA-MUO, na sigla em inglês). Ele é obtido por meio da substituição da série de recompensas (r_k + … + r_m) de uma IA-AR por uma função de utilidade que pode depender de todo o histórico futuro de interações dessa IA:

$$y_k = \arg\max_{y_k} \sum_{x_k y x_{k+1:m}} U(yx_{\leq m}) P(yx_{\leq m} \mid yx_{<k} y_k).$$

Essa fórmula fornece uma maneira de evitar o problema do *wireheading* porque uma função de utilidade definida levando-se em consideração todo o histórico de interações poderia penalizar históricos de interações que apresentassem sinais de autoengano (ou que mostrassem que o agente não estaria investindo recursos suficientes para a obtenção de uma visão apurada da realidade).

Sendo assim, em princípio, com uma IA-MUO seria possível driblar o problema de *wireheading*. Entretanto, para tirar proveito dessa possibilidade, seria necessário especificar uma função de utilidade adequada para a classe de possíveis históricos de interação — uma tarefa que parece proibitivamente difícil.

Talvez fosse mais natural especificar funções de utilidade diretamente em termos de mundos possíveis (ou das propriedades de mundos possíveis, ou de teorias a respeito do mundo), e não de acordo com o histórico de interações de um agente. Se utilizarmos essa abordagem, podemos reformular e simplificar a noção de otimalidade de uma IA-MUO:

$$y = \arg\max_y \sum_w U(w) P(w \mid Ey).$$

Nesse caso, *E* é o total de evidências disponíveis ao agente (no momento em que ele estiver tomando uma decisão) e *U* é a função de utilidade que atribui utilidade a alguma classe de mundos possíveis. O agente ótimo escolhe a ação que maximiza a utilidade prevista.

O problema em aberto nessas formulações é a dificuldade de definir a função de utilidade *U*. Isso nos traz de volta, finalmente, ao problema de inserção de valores. Para fazer com que a função de utilidade seja aprendida pela IA, devemos expandir nosso formalismo para que ele permita algum grau de incerteza sobre funções de utilidade. Isso pode ser feito da seguinte maneira para um agente de inserção de valores (IA-IV):[421]

$$y = \arg\max_{y \in Y} \sum_{w \in W} P(w \mid Ey) \sum_{u \in U} U(w) P(\mathcal{V}(U) \mid w).$$

Aqui, V(.) é uma função que leva funções de utilidade a proposições a respeito de funções de utilidade. V(*U*) é a proposição de que a função de utilidade *U* satisfaz o *critério de valor* expresso por V.[422]

Logo, para decidir qual ação deve ser tomada, o procedimento poderia ser o seguinte: primeiro, calcular a probabilidade condicional de cada mundo possível *w* (com base nas evidências disponíveis e na suposição de que a ação y será tomada). Segundo: para cada função de utilidade possível *U*, calcular a probabilidade condicional de que *U* satisfaça o critério de valor V (com a condição de que *w* seja um mundo possível). Terceiro: para cada função de utilidade possível *U*, calcular a utilidade de mundos possível *w*. Quarto: combinar essas quantidades para calcular a utilidade esperada da ação *y*. Quinto: repetir esse procedimento para cada ação possível e executar a ação para a qual se tenha encontrado a maior utilidade esperada (usando algum método arbitrário de desempate). Conforme descrito, esse procedimento — que envolve considerar cada mundo possível de forma explícita e separadamente — é, logicamente, computacionalmente impraticável. A IA teria que utilizar algum tipo de atalho computacional para fazer uma aproximação dessa noção de otimalidade.

A questão, então, é como definir esse critério de valor V.[423] Uma vez que a IA tivesse uma representação adequada do critério de valor, ela poderia, em princípio, utilizar sua inteligência para coletar informações a respeito de quais mundos possíveis estariam mais próximos da realidade. Ela, então, aplicaria o critério de valor para cada mundo possível *w* que fosse plausível, de modo a encontrar qual função de utilidade satisfaz o critério V em *w*. Podemos, desse modo, considerar a fórmula de uma IA-IV como uma maneira de identificar e isolar o desafio central da abordagem de aprendizado de valores — o desafio da representação. O formalismo também traz à tona uma série de outras questões (por exemplo, como definir Y, W e U) que deveriam ser resolvidas para que a abordagem pudesse funcionar.[424]

Outra questão em relação à maneira de se codificar o objetivo “maximize a realização dos valores descritos dentro do envelope” é que mesmo se todos os valores corretos estivessem descritos em uma carta, e ainda que o sistema de motivação da IA tivesse sido inserido de forma bem-sucedida em seu código, a IA poderia não interpretar essas descrições da forma que pretendíamos. Isso criaria o risco de uma instanciação perversa, conforme discutimos no capítulo 8.

Para esclarecer, a dificuldade aqui não é assegurar que a IA seja capaz de entender as intenções humanas. Uma superinteligência conseguiria fazer isso facilmente. Ao contrário, a dificuldade consiste em assegurar que a IA estará motivada a buscar os valores descritos da maneira que pretendemos. Isso não é algo garantido pela habilidade de compreensão da IA em relação às nossas intenções: uma IA poderia saber exatamente o que estamos querendo dizer e, ainda assim, ser indiferente à interpretação de nossas palavras (em vez disso, ela poderia ser motivada por alguma outra interpretação de nossas palavras ou simplesmente ser totalmente indiferente ao que dizemos).

A dificuldade é agravada pelo desiderato de que, por razões de segurança, a motivação correta deveria, idealmente, ser instalada na IA embrionária *antes* que ela se tornasse capaz de representar por completo conceitos humanos ou pudesse entender intenções humanas. Isso demandaria a criação de uma arquitetura cognitiva que tivesse um lugar específico designado no sistema de motivação da IA para funcionar como um repositório do seu valor final. Mas a arquitetura cognitiva em si teria que admitir revisões, de modo a permitir que a IA expandisse suas capacidades representacionais conforme fosse aprendendo mais a respeito do mundo e se tornando mais inteligente. A IA teria que passar pelo equivalente a revoluções científicas, e então sua visão de mundo seria abalada e ela talvez sofresse crises ontológicas a partir das quais descobriria que o modo como ela pensava anteriormente a respeito de seus valores era baseado em confusões e ilusões. Contudo, desde seu surgimento com um nível sub-humano de desenvolvimento e ao longo do seu subsequente progresso, até se transformar em uma superinteligência galáctica, a conduta da IA precisará ser guiada por um valor final essencialmente imutável e que deverá ser melhor compreendido por ela conforme sua inteligência geral progrida — e cuja compreensão, por parte de uma IA madura, seria bem diferente daquela de seus programadores originais, embora essa diferença pudesse não apresentar um caráter aleatório ou hostil, mas benigno. De que maneira isso seria possível permanece uma questão em aberto.[425] (Ver quadro 11.)

Em resumo, ainda não sabemos como usar a abordagem de aprendizado de valores para instalar valores humanos plausíveis (ver quadro 12 para alguns exemplos de ideias recentes a respeito dessa questão). Até o momento, essa abordagem deve ser vista como uma linha de pesquisa, e não como uma técnica disponível. Se pudéssemos fazê-la funcionar, ela poderia ser a solução mais ideal para o problema de inserção de valores. Entre outros benefícios, essa abordagem parece oferecer uma maneira natural de prevenir crimes da mente, já que uma IA embrionária que fizesse prognósticos razoáveis a respeito de quais valores

poderiam ter sido instalados por seus programadores seria capaz de prever que crimes da mente seriam, provavelmente, avaliados de forma negativa por esses valores e que, portanto, deveriam ser evitados, pelo menos até que informações mais definitivas fossem obtidas.

Por último, mas não menos importante, há ainda outra questão: "o que escrever dentro do envelope" — ou, de modo menos metafórico, quais valores deveríamos tentar ensinar para a IA. Mas essa é uma questão comum a todas as abordagens que tratam do problema da inserção de valores em uma IA. Retornaremos a ela no capítulo 13.

QUADRO 11

Uma IA que deseja ser amigável

Eliezer Yudkowsky tentou descrever algumas características da arquitetura de uma IA embrionária criada para ter o tipo de comportamento descrito acima. De acordo com a terminologia utilizada por ele, essa IA usaria uma "semântica de referência externa".[426] Para ilustrar a ideia básica por trás disso, vamos supor que nosso intuito é que o sistema seja "amigável". O sistema iniciaria com o objetivo de tentar instanciar uma propriedade *F*, embora inicialmente não tenha muitas informações a respeito de *F*. Ele poderia apenas saber que *F* é uma propriedade abstrata e que, quando os programadores falam a respeito de "amizade", eles estão, muito provavelmente, tentando transmitir informações a respeito de *F*. Uma vez que o objetivo final da IA é instanciar *F*, obter mais informações a respeito do que *F* poderia ser teria um importante valor instrumental para a IA. Conforme ela descobrisse mais a respeito de *F*, seu comportamento seria cada vez mais guiado pelo real conteúdo de *F*. Desse modo, talvez a IA se tornasse cada vez mais amigável à medida que fosse aprendendo e ficando mais inteligente.

Os programadores poderiam ajudar nesse processo e reduzir o risco de que a IA cometesse algum erro catastrófico enquanto seu entendimento a respeito de *F* ainda estivesse incompleto. Para isso, eles teriam que disponibilizar para a IA algumas "afirmações dos programadores" — hipóteses em relação à natureza e ao conteúdo de *F* às quais inicialmente a IA atribuiria alta probabilidade de serem verdadeiras. Por exemplo, a hipótese "enganar os programadores é uma atitude hostil" pode receber uma alta probabilidade *a priori*. Essas afirmações dos programadores, entretanto, não são "verdadeiras por definição" — não se tratam de axiomas imutáveis a respeito do conceito de amizade. Pelo contrário, elas são hipóteses iniciais sobre amizade, hipóteses às quais uma IA racional imputaria uma alta probabilidade de serem verdadeiras, pelo menos enquanto ela confiasse mais nas capacidades epistêmicas dos programadores do que nas suas.

A proposta de Yudkowsky também envolve o uso daquilo que chamaremos de "semântica de validade causal". A ideia é que a IA não deveria fazer exatamente aquilo que fosse ordenado pelos programadores, mas (aproximadamente) o que eles estivessem tentando pedir que ela fizesse. Enquanto os programadores estivessem tentando explicar a uma IA embrionária o que é amizade, eles mesmos poderiam cometer alguns erros durante essa explicação. Mais do que isso, os próprios programadores poderiam não compreender por completo a verdadeira natureza da amizade. Seria interessante, então, que a IA tivesse a habilidade de corrigir erros no raciocínio dos programadores e que fosse capaz de inferir o significado verdadeiro ou pretendido de qualquer explicação imperfeita fornecida pelos programadores. Por exemplo, a IA deveria ser capaz de representar os processos causais pelos quais os programadores aprendem e comunicam informações a respeito da amizade. Dessa forma, usando aqui um exemplo trivial, a IA deveria ser capaz de entender que existe a possibilidade de que o programador cometa um erro de digitação ao inserir informações sobre a amizade para então procurar corrigir tal erro. De modo mais geral, a IA deveria buscar corrigir quaisquer distorções que possam ter corrompido o fluxo de informações sobre a amizade desde a fonte, passando pelos programadores até a IA (aqui "distorção" é uma categoria epistêmica). Idealmente, conforme a IA amadurecesse, ela deveria superar quaisquer vieses cognitivos e outras concepções errôneas mais fundamentais que possam ter impedido a total compreensão, por parte de seus programadores, a respeito do que a amizade realmente é.

Modulação de emulação

Em relação ao problema de inserção de valores, podemos encontrar algumas diferenças entre as emulações completas do cérebro e as inteligências artificiais. Métodos que pressupõem um entendimento e controle precisos de algoritmos e

arquitetura não são aplicáveis a emulações. Por outro lado, o método de seleção de motivação por ampliação de capacidade — que não é aplicável a uma inteligência artificial criada do zero — poderia ser usado em emulações (ou cérebros humanos aprimorados).[427]

O método de ampliação de capacidade poderia ser combinado com técnicas de manipulação dos objetivos herdados pelo sistema. Por exemplo, seria possível tentar manipular o estado motivacional de uma emulação administrando o equivalente digital de uma substância psicoativa (ou, no caso de sistemas biológicos, um componente químico real). Mesmo hoje em dia, a farmacologia já é capaz de manipular, ainda que limitadamente, valores e motivações.[428] A farmacopeia do futuro pode conter drogas com efeitos mais específicos e previsíveis. O meio digital das emulações deve facilitar muito tais desenvolvimentos, uma vez que viabilizaria a realização de experimentos controlados e permitiria acesso direto a todas as partes do cérebro.

Semelhante ao caso de animais que são utilizados como cobaias, a pesquisa com emulações teria que lidar com um emaranhado de complicações éticas, muitas das quais não poderiam ser colocadas de lado apenas com a assinatura de um termo de consentimento. Isso poderia atrasar o progresso da tecnologia de emulação (em virtude de regulamentações ou restrições morais), talvez afetando principalmente estudos a respeito da manipulação da estrutura motivacional de emulações. O processo poderia resultar na ampliação das capacidades cognitivas das emulações a níveis potencialmente perigosos de superinteligência antes que testes e ajustes adequados em seus objetivos finais pudessem ser feitos. Outro efeito possível dessas complicações morais seria a liderança do desenvolvimento das tecnologias de emulação por parte de equipes e nações inescrupulosas. Por outro lado, se optássemos por um padrão moral mais flexível em relação aos experimentos com mentes digitais humanas, poderíamos nos tornar responsáveis por uma quantidade substancial de delitos e malefícios — algo obviamente indesejável. Se tudo o mais permanecer constante, essas considerações favorecem a busca por algum caminho alternativo que não demande o uso extensivo de cobaias digitais humanas em situações de alto risco estratégico.

Essa questão, entretanto, não é clara. É possível argumentar que pesquisas em emulação completa do cérebro têm *menos* probabilidade de implicar em violações morais do que pesquisas em inteligência artificial porque provavelmente seríamos mais capazes de identificar quando uma mente emulada estaria qualificada a apresentar status moral do que poderíamos fazê-lo em relação a uma mente completamente desconhecida ou sintética. Se alguns tipos de IA, ou seus subprocessos, tiverem um status moral significativo que não sejamos capazes de reconhecer, as violações morais consequentes poderiam ser extensas. Considere, por exemplo, o grande desprendimento com que os programadores contemporâneos criam agentes de aprendizado por reforço e os sujeitam a estímulos adversos. Inúmeros agentes como esses são criados diariamente, não apenas em laboratórios de ciência da computação, mas em muitos aplicativos, incluindo jogos de computador que contêm personagens autônomos complexos. Presumivelmente, esses agentes ainda são muito primitivos para apresentar

qualquer status moral. Mas até que ponto estamos convencidos de que isso seja realmente verdade? E mais importante do que isso: o quão estamos convencidos de que seremos capazes de parar a tempo, antes que nossos programas se tornem capazes de experimentar qualquer tipo de sofrimento moralmente relevante?

(No capítulo 14 retomaremos algumas das questões estratégicas mais amplas que surgem quando ponderamos a respeito de qual seria o melhor caminho a seguir: o da emulação ou a inteligência artificial.)

QUADRO 12

Duas ideias recentes (e pouco amadurecidas)

Uma abordagem que poderíamos chamar de "ave-maria" é baseada na esperança de que, em algum outro lugar do universo, existam (ou virão a existir) civilizações bem-sucedidas em controlar uma explosão de inteligência e com valores semelhantes aos nossos. Nós poderíamos, então, tentar construir nossa IA de modo que ela seja motivada a fazer aquilo que outras superinteligências gostariam que ela fizesse.[429] A vantagem aqui é que essa tarefa pode ser mais simples do que construir nossa IA de maneira que ela fosse motivada a fazer diretamente o que desejamos.

Para que esse esquema funcionasse, não seria necessário que nossa IA conseguisse estabelecer uma comunicação direta com uma superinteligência alienígena. Em vez disso, as ações de nossa IA seriam guiadas pelas *suas estimativas* a respeito daquilo que uma IA alienígena desejaria que ela fizesse. Nossa IA criaria modelos para os possíveis resultados de uma explosão de inteligência em outros lugares e, conforme ela mesma se tornasse superinteligente, suas estimativas deveriam se tornar cada vez mais precisas. Não é necessário um conhecimento perfeito. Pode existir uma gama de resultados plausíveis para explosões de inteligência, e nossa IA faria o seu melhor para acomodar em suas estimativas as preferências dos diferentes tipos de superinteligência que poderiam surgir, ponderadas pela probabilidade de cada tipo.

Para essa versão da abordagem "ave-maria", teríamos que construir um valor final para a nossa IA que fizesse referência às preferências de outras superinteligências. Como exatamente isso poderia ser feito ainda não é algo muito claro. Contudo, agentes superinteligentes poderiam ser estruturalmente distintos o suficiente para que fôssemos capazes de escrever um programa de computador que funcionaria como um detector que olhasse para o modelo de mundo em nossa IA em desenvolvimento e designasse os elementos representacionais correspondentes à presença de uma superinteligência. Esse detector poderia, então, de alguma maneira, extrair as preferências da superinteligência em questão (da maneira como ela fosse representada por nossa própria IA).[430] Se fôssemos capazes de criar tal detector, poderíamos usá-lo para definir os valores finais de nossa IA. Um desafio é que precisaríamos criar o detector antes de termos conhecimento a respeito de qual estrutura representacional nossa IA desenvolveria. Desse modo, o detector talvez precisasse consultar uma estrutura representacional desconhecida e extrair as preferências de qualquer superinteligência que estivesse representada dentro dessa estrutura. Essa parece ser uma tarefa difícil, mas talvez alguma solução perspicaz possa ser encontrada.[431]

Se essa configuração básica funcionasse, uma série de melhorias seria imediatamente óbvia. Por exemplo, em vez de tentar seguir (alguma composição ponderada de) preferências de *toda e qualquer* superinteligência alienígena, o valor final de nossa IA poderia ter um filtro que selecionasse um subgrupo de superinteligências que seriam seguidas (com o objetivo de selecionar aquelas que tivessem valores semelhantes aos nossos). Por exemplo, poderíamos usar critérios relativos à origem causal de uma superinteligência para determinar se ela deveria ou não fazer parte desse subgrupo. Algumas propriedades relacionadas à origem dessa IA (que talvez possamos definir em termos estruturais) poderiam ter correlação com o grau de proximidade que seus valores teriam com os nossos. Talvez devêssemos confiar mais em superinteligências que possuam origens causais em emulações completas do cérebro, ou em uma IA embrionária que não tenha feito uso intensivo de algoritmos evolutivos ou que tenha emergido de forma lenta — o que poderia sugerir uma partida controlada. (Levar as origens causais em consideração também evitaria que atribuíssemos muito peso a superinteligências que criassem muitas cópias de si mesma — inclusive poderíamos evitar dar um incentivo para que elas fizessem isso.) Diversas outras melhorias também seriam possíveis.

A abordagem "ave-maria" requer que tenhamos fé na existência de outras superinteligências no universo afora com valores suficientemente semelhantes aos nossos.[432] Isso faz com que essa não seja uma abordagem ideal. No entanto, embora os obstáculos técnicos em torno da abordagem "ave-maria" sejam bastante substanciais, eles são, possivelmente, menos formidáveis do que aqueles que seriam enfrentados caso optássemos por abordagens alternativas. Explorar abordagens não ideais, porém mais simples de serem

implementadas, faz algum sentido — não com o objetivo de usá-las, mas para ter algo que pudesse ser utilizado caso uma solução ideal não esteja pronta a tempo.

Outra ideia de como resolver o problema de inserção de valores foi proposta por Paul Christiano.[433] Assim como a "ave-maria", essa abordagem é um método de aprendizado de valores que procura definir o critério de valor através de um "truque" em vez de fazê-lo por meio de uma construção laboriosa. Mas, ao contrário da "ave-maria", ela não pressupõe a existência de outros agentes superinteligentes que serviriam como modelo para a nossa IA. A proposta de Christiano resiste a uma explicação breve — ela envolve uma série de considerações enigmáticas —, mas podemos fazer, pelo menos, uma tentativa de descrever seus principais elementos.

Suponha que possamos obter (a) uma especificação matemática precisa de um cérebro humano específico e (b) um ambiente virtual matematicamente bem especificado que contenha um computador idealizado com uma quantidade arbitrariamente grande de poder de processamento e memória. Dados (a) e (b), poderíamos definir uma função de utilidade *U* como a saída de dados que o cérebro humano produziria após interagir com esse ambiente. *U* seria um objeto matematicamente bem definido, embora talvez não seja possível descrevê-lo *explicitamente* (em virtude de limitações computacionais). De qualquer modo, *U* poderia servir como o critério de valor para uma IA com aprendizado de valores, que poderia utilizar heurísticas variadas para atribuir probabilidades a hipóteses sobre o que *U* sugere.

Intuitivamente, gostaríamos que *U* representasse a função de utilidade que uma humana devidamente preparada poderia gerar se tivesse a vantagem de ser capaz de usar uma quantidade arbitrariamente grande de poder computacional — poder computacional suficiente para, por exemplo, rodar números astronômicos de cópias dela mesma que fossem capazes de auxiliá-la em sua análise na busca por uma especificação de uma função de utilidade ou ajudá-la a criar um processo mais eficiente para realizar essa análise. (Nós estamos aqui antevendo um tema, "vontades extrapoladas coerentes", que será explorado mais detalhadamente no capítulo 13.)

Especificar um ambiente idealizado pode parecer algo relativamente fácil: podemos fornecer uma descrição matemática de um computador abstrato com capacidade arbitrariamente grande; de outro modo, poderíamos utilizar um programa de realidade virtual que fornecesse uma descrição matemática de, digamos, uma sala com um terminal de computador em seu interior (instanciando o computador abstrato). Mas como obter uma descrição matematicamente precisa de um cérebro humano específico? A maneira mais óbvia seria através de uma emulação completa do cérebro, mas e se a tecnologia necessária para uma emulação não estiver disponível a tempo?

É nesse ponto que a proposta de Christiano apresenta uma inovação importante. Ele observa que, para obter um critério de valor matematicamente bem especificado, não é necessário que tenhamos um modelo computacional funcional de uma mente, um modelo que pudéssemos rodar. Seria necessário apenas que tivéssemos uma *definição* matemática (possivelmente implícita e extremamente complicada) — e obter isso pode ser muito mais simples. Utilizando neuroimageamento funcional e outras técnicas de medição, talvez conseguíssemos coletar gigabytes de informações a respeito do padrão de entrada e saída de dados em um humano selecionado. Se coletarmos uma quantidade suficiente de dados, talvez o modelo matemático mais simples que conseguisse explicar todos esses elementos seria, de fato, uma emulação desse humano específico em questão. Embora *encontrar* esse modelo mais simples a partir dos dados levantados seja algo intratável computacionalmente, poderia ser perfeitamente possível *definir* esse modelo se fizermos referência aos dados e nos utilizarmos de uma medida de simplicidade matematicamente bem definida (como alguma variante da complexidade de Kolmogorov, que vimos no quadro 1, capítulo 1).[434]

Projeto institucional

Alguns sistemas inteligentes se constituem de partes inteligentes que são, elas mesmas, capazes de atuar de forma autônoma. No mundo humano, firmas e nações são exemplos desse tipo de sistema: embora sejam formadas por seres humanos, elas podem, para alguns propósitos, ser vistas como agentes autônomos independentes. As motivações de tais sistemas compostos dependem não apenas das motivações dos subagentes que os constituem, mas também de como esses subagentes estão organizados. Por exemplo, um grupo que esteja organizado sob uma forte ditadura pode se comportar como se tivesse vontades idênticas àquelas do subagente que ocupa a posição de ditador, enquanto um grupo democrático pode às vezes se comportar como se suas vontades fossem uma composição ou a

média das vontades de seus vários constituintes. Mas também podemos imaginar instituições de governança que fariam com que uma organização se comportasse de um modo que não pudesse ser representado por uma função simples das vontades de seus subagentes. (Ao menos teoricamente, seria possível supor a existência de um Estado totalitário que fosse odiado por *todos* porque o Estado teria mecanismos que preveniriam que seus cidadãos coordenassem uma revolta. Cada cidadão poderia ficar numa situação de desvantagem caso decidisse se revoltar por conta própria em vez de fazer sua parte na máquina estatal.)

Desse modo, por meio da criação de projetos de instituições apropriadas para um sistema composto, poderíamos tentar moldar sua motivação efetiva. No capítulo 9, discutimos que a integração social poderia funcionar como um método possível de controle de capacidade. Mas nesse capítulo estivemos focados nos incentivos encontrados por um agente como consequência de sua existência em um mundo social onde todos seriam praticamente iguais. Aqui, nos concentraremos no que poderia acontecer *dentro* de um determinado agente: de que maneira sua vontade é determinada por sua organização interna. Estamos, desse modo, analisando um método de seleção de motivação. Além disso, como esse tipo de projeto institucional interno não depende de uma engenharia ou reforma social em larga escala, trata-se de um método que poderia estar disponível para um projeto individual cujo objetivo seria desenvolver uma superinteligência, mesmo que o meio socioeconômico ou internacional não seja idealmente favorável.

O projeto institucional talvez seja mais plausível em contextos nos quais ele fosse combinado com técnicas de ampliação de capacidade. Se pudéssemos iniciar com agentes que já estivessem adequadamente motivados ou que já tivessem motivações parecidas com as humanas, arranjos institucionais poderiam ser usados como uma salvaguarda adicional para aumentar as chances de manter o sistema no curso certo.

Por exemplo, suponha que iniciássemos com agentes semelhantes a humanos bem motivados — como as emulações, por exemplo. Gostaríamos de aumentar as capacidades cognitivas desses agentes, mas existiria a preocupação de que essas melhorias cognitivas pudessem corromper suas motivações. Uma forma de lidar com esse desafio seria por meio da configuração de um sistema em que emulações individuais funcionassem como subagentes. Quando uma certa melhoria fosse introduzida, ela seria aplicada primeiramente em um pequeno grupo de subagentes. Então, seus efeitos poderiam ser estudados por um grupo de revisão composto de subagentes que ainda não possuíssem tal melhoria. Essas melhorias seriam aplicadas no restante da população de subagentes apenas quando esse segundo grupo estivesse certo de que elas não corromperam as motivações do grupo que as recebeu. Se fosse verificado que os subagentes que receberam as melhorias tiveram suas motivações corrompidas, eles não receberiam nenhuma melhoria posterior e seriam excluídos de posições centrais em processos de tomada de decisão (pelo menos até o ponto em que o sistema como um todo estivesse suficientemente evoluído para que os subagentes corrompidos pudessem ser reintegrados de maneira segura).[435] Embora os subagentes corrompidos pudessem ganhar algum tipo de vantagem por conta das melhorias que receberiam,

a estrutura institucional em que estariam inseridos e o fato de que constituiriam uma pequena minoria de subagentes impediriam (assim esperamos) que eles tomassem o poder ou corrompessem o restante do sistema. Desse modo, a inteligência coletiva e a capacidade do sistema poderiam ser gradualmente ampliadas em uma sequência de pequenos passos, onde a solidez de cada passo seria verificada por subagentes apenas um pouco menos capacitados do que os novos subagentes que estivessem sendo introduzidos nessa etapa.

A quantidade de segurança passível de ser obtida por meio de um projeto institucional como esse é limitada pela precisão dos testes que seriam aplicados para avaliar os subagentes que recebessem melhorias. O problema em torno de sua elaboração é agravado pelo fato de que as baterias de testes poderiam ser validadas apenas com base em uma faixa de habilidades cognitivas preexistentes, e esses testes seriam aplicados a agentes que se encontrariam fora dessa faixa. Esse problema poderia ser atenuado se as melhorias fossem feitas em pequenos passos e se os testes fossem executados por um longo período de tempo. Tal precaução, entretanto, aumentaria os custos e diminuiria a velocidade dos avanços (e caso uma corrida tecnológica estivesse ocorrendo, isso poderia significar que um projeto que empregasse tais medidas de precaução estaria se colocando em desvantagem). Mesmo com pequenos passos e testes extensivos, ainda restaria algum risco de que falhas sutis fossem introduzidas, o que poderia levar a uma corrupção gradual do sistema de objetivos, a qual poderia se manifestar apenas quando o sistema alcançasse a superinteligência. Esse risco poderia ser reduzido se deixássemos que esses estágios de teste se estendessem por diversas etapas do processo de aprimoramento cognitivo. Desse modo, mesmo os agentes inicialmente classificados como seguros e que tenham recebido diversas melhorias subsequentes continuariam sendo avaliados por um grupo de agentes “originais” inalterados e não corrompidos e estariam sujeitos a reversão.

Idealmente, em vez de ocorrer apenas num estágio inicial de avaliação, esses testes deveriam ser contínuos. Subagentes de nível humano teriam a habilidade de criar estratégias e poderiam, dessa forma, optar por esconder certos objetivos enquanto seu comportamento estivesse sendo monitorado. Para que testes contínuos pudessem ser realizados, talvez fosse necessária a criação de um tipo de hierarquia na qual os subagentes com um certo nível de capacidade fossem encarregados do monitoramento dos subagentes que tivessem um nível um pouco maior de capacidade. Na parte inferior dessa hierarquia de capacidade (mas no topo da hierarquia de poder) estaria o relativamente estúpido e lento diretor. Esse diretor humano seria como um rei demente que reina sobre uma corte incompetente e supervisiona uma administração medíocre, mas que governa um povo competente. Pensado de outro modo, esse diretor humano poderia ser comparado ao “cérebro reptiliano”, evolucionariamente primitivo, que assegura que o neocórtex, mais novo e muito mais capacitado, seja empregado a serviço de objetivos como alimentação e copulação.

Podemos nos perguntar se uma meritocracia inversa como essa poderia se manter estável. Se a maioria dos subagentes superinteligentes estivessem corrompidos e os humanos responsáveis por sua operação tentassem desfazer

muitos dos procedimentos de melhoria anteriores, a estrutura institucional seria forte o suficiente para resistir à oposição que tais mudanças gerariam? No contexto das sociedades humanas contemporâneas ou ao longo da história, isso se mostra como algo implausível. (Não é sempre que as pessoas mais capacitadas conseguem alcançar as posições de poder, mas também não somos *literalmente* governados por chimpanzés.) Entretanto, essa questão é bem menos clara dentro de um contexto no qual a sociedade tenha sido implementada como um software.[436]

Imagine uma hierarquia digital que possua um agente policial altamente treinado para cada dois proletários. E, além disso, um agente policial especial para cada dois policiais comuns. E, além disso, um agente policial especial-especial para cada dois agentes policiais especiais — e assim por diante, de modo que cada camada de vigilantes tenha pelo menos metade do número de indivíduos da camada que é vigiada por ela. Indo mais além, imagine que essa sociedade seja arquitetada de tal modo que um supervisor tenha enormes vantagens em relação a seus subordinados. Por exemplo, o supervisor seria capaz de monitorar e gravar tudo o que seus subordinados fizessem ou dissessem, mantendo-se escondido atrás de um espelho, de maneira que seus subordinados pudessem apenas ouvir as ordens e instruções que ele escolhesse transmitir. Imagine que esses supervisores teriam à disposição um painel de botões que poderiam ser acionados a qualquer momento para punir, recompensar, pausar, apagar ou reverter seus subordinados a um estágio anterior. Essas capacidades aumentariam em muito os poderes de um agente supervisor para além de qualquer coisa que já tenha sido possível historicamente, mesmo na mais totalitária das organizações.

Mas isso não esgota os possíveis instrumentos de controle. Os supervisores poderiam ter também ao seu dispor simulações de realidade virtual realísticas que seriam utilizadas para sondar as reações de seus subordinados diante de cenários hipotéticos (tais como a possibilidade de aderir a uma conspiração ou deixar de trabalhar caso pensassem que ninguém os estivesse observando). Além disso, os supervisores também poderiam ter a habilidade de espionar até mesmo as mentes de seus subordinados para manipular diretamente seus estados emocionais.

O resultado seria uma série de vantagens expressivas para o supervisor.[437] Cada supervisor seria subordinado a outro supervisor que se encontraria um nível acima e que monitoraria cada passo *seu* e se certificaria de que *ele* não deixaria de cumprir à risca suas tarefas de gerenciamento. Com esse tipo de arranjo, provavelmente teríamos uma estrutura suficientemente estável, ainda que cada nível de supervisão fosse intelectualmente inferior ao nível que estivesse sob sua vigilância.[438] E toda a estrutura, que incluiria muitos agentes altamente superinteligentes, poderia ser controlada por um pequeno grupo de humanos que tivesse acesso com permissões de administrador.[439]

O custo de tal estrutura interna de controle seria significativo, mas não necessariamente proibitivo. Por exemplo, no modelo em que cada nível de supervisão tivesse metade do número de membros que compõem o nível inferior, o gasto computacional adicional se limitaria a 100% em relação ao que a parte proletária do sistema custaria — e seria ainda menor se a camada dos administradores humanos menos inteligentes demandasse menos computações por

subagente. A fiscalização seria ainda maior se aumentássemos a proporção de supervisores por subagente ou o custo poderia ser amenizado caso diminuíssemos essa proporção. O custo computacional extra seria um fator primordial em um cenário competitivo (cf. capítulo 11), mas parece razoável dentro de um contexto que envolvesse um projeto que não estivesse enfrentando uma forte competição imediata. Existiria também um custo em termos do tempo de desenvolvimento necessário para a criação e teste dessas funcionalidades de supervisão. Um projeto bem financiado poderia reduzir esse custo de tempo se desenvolvesse a estrutura de controle e a inteligência de máquina em paralelo, mas isso demandaria uma quantidade extra de recursos que seria proibitiva para projetos menores ou que se encontrassem em meio a uma disputada corrida tecnológica.

Existe um outro tipo de custo que também merece ser considerado: o risco de que crimes da mente sejam cometidos dentro desse tipo de estrutura.[440] Segundo nossa descrição, essa instituição se assemelharia a um horrível campo de trabalho forçado na Coreia do Norte. No entanto, existem maneiras de pelo menos mitigar os problemas morais que envolvem a administração desse tipo de instituição, ainda que os subagentes que a compõem sejam emulações com status moral igual ao humano. Na pior das hipóteses, o sistema poderia contar apenas com emulações que participassem voluntariamente. Então, cada subagente teria a opção de se retirar da instituição em qualquer momento.[441] Emulações que tivessem sido finalizadas poderiam ser armazenadas na memória com o compromisso de que seriam reiniciadas sob condições muito melhores, quando ultrapassada a perigosa fase de explosão de inteligência. Nesse meio-tempo, os subagentes que optassem por participar seriam instalados em confortáveis ambientes virtuais e poderiam ter garantidos amplos períodos de descanso e lazer. Essas medidas demandariam um custo, que seria viável para um projeto bem financiado e que não estivesse em um cenário competitivo. Em uma situação de alta competitividade, esse custo poderia ser inviável, a não ser que todos os projetos competidores incorressem, necessariamente, nos mesmos custos.

No exemplo acima, imaginamos os subagentes como emulações. Assim, poderia surgir um questionamento: a abordagem de projeto institucional requer que os subagentes sejam antropomórficos? Ou ela seria igualmente aplicável a sistemas compostos por subagentes artificiais?

O primeiro pensamento que surgiria a esse respeito pode ser um tanto cético. Notamos que, apesar de nossa larga experiência com agentes do tipo humano, ainda não somos capazes de prever a eclosão ou resultados de revoluções. As ciências sociais podem, quando muito, descrever algumas tendências estatísticas.[442] Uma vez que não podemos prever de forma confiável a estabilidade de estruturas sociais no caso de seres humanos comuns (a respeito dos quais possuímos muita informação), é tentador inferir que existe pouca esperança de que sejamos capazes de arquitetar minuciosamente estruturas sociais estáveis para agentes do tipo humano cognitivamente melhorados (sobre os quais não possuímos informação alguma) e menos esperança ainda de conseguir fazê-lo no caso de agentes artificiais avançados (que não são nem ao menos similares aos agentes sobre os quais possuímos alguma informação).

No entanto, essa questão não é tão óbvia. Humanos e seres semelhantes aos humanos são complexos, mas agentes artificiais poderiam apresentar arquiteturas relativamente simples. Os agentes artificiais poderiam inclusive ter motivações simples e explícitas. Além do mais, agentes digitais (quer sejam emulações ou inteligências artificiais) são geralmente copiáveis: essa é uma possibilidade que poderia revolucionar o gerenciamento do mesmo modo que a substituição de peças revolucionou a indústria. Essas diferenças, aliadas à oportunidade de trabalhar com agentes que inicialmente não seriam poderosos e à criação de estruturas institucionais que utilizem várias das já mencionadas medidas de controle, poderiam ser combinadas, tornando possível a obtenção de um tipo particular de resultado institucional — um sistema que não se revolte — que seria mais confiável do que os sistemas formados por seres humanos sob condições históricas.

Mas, novamente, agentes artificiais podem não ter muitos dos atributos existentes em agentes do tipo humano que nos ajudam a prever seu comportamento. Agentes artificiais não necessitam de nenhum tipo de emoção social como aquelas que moldam os comportamentos humanos, emoções como medo, orgulho e remorso. Os agentes artificiais também não precisam criar nenhum vínculo que os ligue à família ou aos amigos. Eles também não precisam exibir a linguagem corporal inconsciente que nos impede de esconder nossas intenções. Esses déficits podem desestabilizar instituições de agentes artificiais. Mais do que isso, agentes artificiais podem ser capazes de dar grandes saltos em desempenho cognitivo como resultado de mudanças aparentemente pequenas em seus algoritmos ou arquitetura. Agentes friamente otimizadores poderiam estar dispostos a correr riscos que fariam qualquer humano tremer.[443] E agentes superinteligentes poderiam apresentar uma surpreendente habilidade de coordenação com pouca ou nenhuma comunicação (criando, por exemplo, modelos internos das respostas hipotéticas uns dos outros diante de diversas contingências). Essas e outras diferenças poderiam tornar mais provável uma falha institucional repentina, mesmo com a presença de métodos de controle social que pareçam infalíveis.

Desse modo, é incerto o quão promissora seria a abordagem de projeto institucional e se ela teria uma chance maior de funcionar com agentes antropomórficos ou com agentes artificiais. Podemos pensar que a criação de uma instituição com diversos níveis de verificação necessariamente aumentaria a segurança — ou, pelo menos, não diminuiria a segurança —, de modo que, a partir de um ponto de vista de mitigação de riscos, seria melhor que o método fosse utilizado. Mas não temos como afirmá-lo isso com toda a certeza. Uma vez que a abordagem adiciona componentes novos e complexidade, ela poderia introduzir novas possibilidades de falha que não existiriam caso o agente não possuísse subagentes inteligentes como componentes. No entanto, a abordagem de projeto institucional merece ser melhor explorada.[444]

TABELA 12

Sumário de técnicas de inserção de valores

Representação explícita	Pode ser promissora como uma maneira de inserir valores de domesticidade. Não parece promissora para a inserção de valores mais complexos.
Seleção evolucionária	Menos promissora. Processos poderosos de busca podem encontrar um design que satisfaça critérios formais de busca, mas não nossas reais intenções. Além disso, se tivermos que executar esses designs para poder avaliá-los — incluindo aqui designs que nem ao menos preenchem os critérios formais —, um grave perigo adicional pode ser criado. A evolução também torna difícil evitar grandes crimes da mente, especialmente se nossa meta for a criação de mentes semelhantes às humanas.
Aprendizagem por reforço	Uma gama variada de métodos pode ser utilizada para resolver "problemas de aprendizagem por reforço", mas eles tipicamente envolvem a criação de um sistema que busca maximizar um sinal de recompensa. Esse tipo de método tem uma tendência inerente de produzir um modo de falha do tipo *wireheading* quando o sistema se torna mais inteligente. Desse modo, a aprendizagem por reforço parece pouco promissora.
Acréscimo de valores	Nós, humanos, adquirimos grande parte de nosso conteúdo específico de valores a partir de nossas reações ao que experimentamos. Embora o acréscimo de valores pudesse, em princípio, ser usado para criar um agente com motivações humanas, a disposição para o acréscimo de valores humanos poderia ser complexa e difícil de replicar em uma IA embrionária. Caso esse acréscimo de valores fosse malsucedido, o resultado poderia ser uma IA que fizesse generalizações de forma distinta dos humanos, o que resultaria na aquisição de valores finais indesejados. Mais pesquisas seriam necessárias para determinar o quão difícil seria fazer com que o acréscimo de valores funcionasse de uma maneira suficientemente precisa.
Estepe motivacional	É muito cedo para afirmar o quão difícil seria conseguir encorajar um sistema a desenvolver representações internas de alto nível que sejam transparentes aos humanos (ao mesmo tempo que mantivesse as capacidades do sistema abaixo de um nível perigoso) e então usar essas representações para criar um novo sistema de objetivos. Essa abordagem aparenta ser bastante promissora. (No entanto, assim como qualquer outra abordagem que ainda não foi testada e que adiaria em muito os esforços para a criação de medidas de segurança até o desenvolvimento de uma IA de nível humano, seria necessário ter cuidado para não permitir que essa abordagem fosse usada como uma desculpa para uma atitude apática em relação ao problema do controle nesse meio-tempo.)
Aprendizado e valores	Uma abordagem potencialmente promissora, embora mais pesquisas sejam necessárias para determinar o quão difícil seria especificar formalmente uma referência que conseguisse apontar para as informações externas relevantes a respeito de valores humanos (e o quão difícil seria especificar um critério de exatidão, em relação a essa referência, em uma função de utilidade). Dentro da categoria de aprendizado de valores, também valeria a pena explorar a proposta de uma abordagem do tipo "ave-maria" ou uma que seguisse a linha proposta por Paul Christiano (ou algum outro atalho semelhante).
Modulação de emulações	Se a inteligência de máquina fosse alcançada por meio de emulações, seria possível manipular suas motivações por meio de equivalentes digitais de drogas ou por algum outro meio. Ainda não sabemos se isso permitiria a inserção suficientemente precisa de valores de modo a garantir segurança mesmo que a emulação atinja a superinteligência. (Restrições éticas também poderiam complicar desenvolvimentos que seguissem essa direção.)
Projeto institucional	Diversos métodos poderosos de controle social poderiam ser aplicados em uma instituição formada por emulações. Em princípio, métodos de controle social também poderiam ser aplicados em instituições formadas por inteligências artificiais. Emulações apresentam algumas propriedades que possibilitariam controlá-las mais facilmente através desses métodos, embora também possuam algumas propriedades que as tornam mais difíceis de

controlar do que as IAs. Aparentemente, a abordagem de projeto institucional merece ser melhor explorada como uma possível técnica de inserção de valores.

Sinopse

A engenharia de sistemas de objetivos ainda não é uma disciplina estabelecida. Ainda não sabemos como transferir valores humanos para um computador digital, mesmo se tratando de máquinas com nível humano de inteligência. Com a investigação de uma série de abordagens, chegamos à conclusão de que algumas delas parecem levar a becos sem saída, mas outras parecem promissoras e merecem ser melhor exploradas. A tabela 12 apresenta um sumário dessas abordagens.

Se soubéssemos como resolver o problema de inserção de valores, estaríamos diante de um outro problema: decidir quais valores inserir. Em outras palavras: o que desejamos que a superinteligência queira? Esse é o problema de cunho mais filosófico e o abordaremos a seguir.

13

CAPÍTULO 13

ELEGENDO OS CRITÉRIOS DE ESCOLHA

Suponha que pudéssemos implantar qualquer valor final arbitrário em uma IA embrionária. A decisão a respeito de qual seria o valor instalado poderia, então, ter consequências de longo alcance. Outras determinadas escolhas de parâmetros básicos – relativas aos axiomas da teoria de decisão e epistemologia da IA – poderiam ser similarmente importantes. Mas tolos, ignorantes e limitados que somos, como poderíamos ser confiáveis para tomar boas decisões em relação ao projeto? Como poderíamos escolher sem deixar impressos para sempre os preconceitos da geração atual? Neste capítulo, exploraremos como a normatividade indireta pode nos permitir repassar à própria superinteligência grande parte do trabalho cognitivo associado à tomada dessas decisões e ainda assim manter o resultado ancorado em valores humanos mais profundos.

A necessidade da normatividade indireta

De que maneira podemos fazer com que uma superinteligência realize o que desejamos? O que queremos que a superinteligência queira? Até este momento, focamos na primeira questão. Agora abordaremos a segunda.

Suponha que tenhamos resolvido o problema do controle de modo que fôssemos capazes de inserir qualquer valor escolhido no sistema de motivação de uma superinteligência, fazendo com que ela buscasse esse valor como seu objetivo final. Que valor deveríamos inserir? Essa não é uma escolha simples. Se a superinteligência viesse a obter uma vantagem estratégica decisiva, esse valor determinaria o destino da herança cósmica.

Fica nítido o quão essencial é não nos equivocarmos em nossa seleção de valores. Mas como poderíamos, realisticamente, atingir infalibilidade em uma questão como essa? Talvez estejamos errados a respeito do que seja a moralidade; e igualmente equivocados sobre o que é bom para nós; e inclusive em relação ao que realmente queremos. Especificar um objetivo final parece demandar o enfrentamento de problemas filosóficos complicados. Se tentarmos uma abordagem direta, é bem provável que criemos alguma confusão. O risco de uma escolha errada é especialmente alto quando o contexto dessa decisão não nos é familiar — e selecionar o objetivo final para uma superinteligência de máquina, que moldará todo o futuro da humanidade, é um contexto de decisão extremamente desconhecido, mais do que qualquer outro.

As chances desanimadoras de sucesso em uma abordagem direta se refletem nas divergências generalizadas em relação às questões relevantes da teoria de valores. Nenhuma teoria ética tem o apoio majoritário dos filósofos, então a maioria dos filósofos deve estar errada.[445] Também estão refletidas nas evidentes

mudanças que a distribuição de crenças morais tem sofrido ao longo do tempo, muitas das quais gostamos de julgar como progressos. Na Europa medieval, por exemplo, era considerado um entretenimento respeitável assistir a um prisioneiro político ser torturado até a morte. A queima de gatos ainda era popular no século XVI em Paris.[446] Há apenas 150 anos, a escravidão ainda era praticada amplamente no Sul dos Estados Unidos, plenamente amparada pelas leis e pelos costumes morais. Quando olhamos para trás, vemos deficiências gritantes não apenas nos comportamentos, mas também nas crenças morais de todas as épocas anteriores. Embora seja provável que tenhamos adquirido algum discernimento moral desde então, dificilmente poderíamos afirmar que desfrutamos hoje de um esclarecimento moral pleno e perfeito. É bem provável que ainda estejamos operando sob um ou mais equívocos morais graves. Em tais circunstâncias, escolher um valor final baseado nas nossas convicções atuais, de forma que ele ficasse cristalizado para sempre sem qualquer possibilidade de um progresso ético mais adiante, seria arriscar uma calamidade moral existencial.

Ainda que pudéssemos confiar racionalmente na identificação de uma teoria ética correta — algo impossível —, continuaríamos correndo o risco de cometer erros durante o desenvolvimento de detalhes importantes dessa teoria. Teorias morais aparentemente simples podem esconder muitas complexidades.[447] Por exemplo, considere a (excepcionalmente simples) teoria consequencialista do hedonismo. Em resumo, essa teoria afirma que exclusivamente o prazer possui valor e que exclusivamente a dor não tem valor.[448] Mesmo se colocássemos todas as nossas apostas morais nessa teoria e ela se verificasse correta, uma gama variada de perguntas continuaria sem resposta. Os “prazeres superiores” deveriam ser priorizados em relação aos “prazeres inferiores”, como argumentou John Stuart Mill? De que maneira a intensidade e a duração do prazer deveriam ser levadas em consideração? As dores e os prazeres poderiam anular um ao outro? Que tipos de estados cerebrais estão associados aos prazeres moralmente relevantes? Duas cópias exatas do mesmo estado cerebral corresponderiam ao dobro da quantidade de prazer?[449] Podem existir prazeres subconscientes? Como deveríamos lidar com chances extremamente pequenas de obter prazeres extremamente intensos?[450] Como isso poderia ser agregado a uma população infinita?[451]

Fornecer uma resposta errada para qualquer uma dessas perguntas poderia ser catastrófico. Se, ao escolher um valor final para a superinteligência, nós tivéssemos que apostar não apenas em uma teoria moral geral, mas em uma conjunção longa de afirmações específicas a respeito de como essa teoria deveria ser interpretada e integrada em um processo efetivo de tomada de decisões, então nossa chance de sucesso diminuiria para algo próximo de zero. Alguns tolos talvez aceitassem impulsivamente esse desafio de resolver, em uma única tacada, todos os problemas mais relevantes da filosofia moral a fim de fixar suas respostas favoritas em uma IA embrionária. Almas mais sábias procurariam cuidadosamente por uma abordagem alternativa, buscando alguma maneira de se proteger.

Isso nos traz à normatividade indireta. A razão óbvia para se construir uma superinteligência é que nós poderíamos deixar a cargo dela o raciocínio instrumental necessário para encontrar maneiras efetivas de se alcançar um dado

valor. A normatividade indireta nos permitiria também deixar por conta da superinteligência alguns dos raciocínios necessários para selecionar o valor que deverá ser atingido.

A normatividade indireta é uma maneira de responder ao desafio apresentado pelo fato de que podemos não saber o que realmente queremos, o que seria de nosso interesse, ou o que é moralmente correto ou ideal. Em vez de tentar adivinhar essas questões com base em nosso conhecimento atual (que tem grandes chances de estar incorreto), nós delegaríamos a uma superinteligência algumas das tarefas cognitivas exigidas para a seleção de valores. Uma vez que a superinteligência seria cognitivamente superior a nós, é provável que ela conseguisse enxergar além dos erros e confusões que ofuscam nosso pensamento. Poderíamos generalizar essa ideia e cunhá-la como um princípio heurístico:

> **O princípio da deferência epistêmica**
> Uma superinteligência futura ocupa uma posição estratégica epistemicamente superior: em comparação às nossas, suas convicções (provavelmente, na maioria dos tópicos) teriam maior possibilidade de ser verdadeiras. Nós deveríamos, portanto, optar pela opinião da superinteligência sempre que possível.[452]

A normatividade indireta aplica esse princípio ao problema de seleção de valores. Em virtude da pouca confiança que temos em nossa habilidade de especificar um padrão normativo concreto, nós especificaríamos, em vez disso, algum tipo mais abstrato de condição que pudesse ser satisfeita por qualquer padrão normativo, na esperança de que uma superinteligência fosse capaz de encontrar um padrão concreto que satisfizesse essa condição abstrata. Nós poderíamos fornecer a essa IA embrionária o objetivo final de atuar continuamente de acordo com a sua melhor estimativa em relação ao que esse padrão implicitamente definido determinaria que ela fizesse.

Alguns exemplos servirão para esclarecer mais essa ideia. Primeiro, consideraremos a “vontade extrapolada coerente”, uma proposta de normatividade indireta delineada por Eliezer Yudkowsky. Depois, apresentaremos algumas variações e alternativas que nos fornecerão uma noção da gama de opções disponíveis.

Vontade extrapolada coerente

Yudkowsky propôs que fosse dado a uma IA embrionária o objetivo final de realizar a “Vontade Extrapolada Coerente” (VEC) da humanidade, a qual ele define da seguinte maneira:

> Nossa vontade extrapolada coerente corresponderia ao nosso desejo caso tivéssemos mais conhecimento, pensássemos mais rápido, fôssemos mais parecidos com as pessoas que gostaríamos de ser e tivéssemos nos desenvolvido mais conjuntamente; onde a extrapolação convergisse em vez de

> divergir, onde nossos desejos fossem coerentes em vez de interferir uns nos outros, extrapolados da forma como gostaríamos que eles fossem extrapolados, interpretados da forma como gostaríamos que eles fossem interpretados.[453]

Quando Yudkowsky escreveu isso, ele não pretendia apresentar um plano de implementação dessa prescrição um tanto poética. Seu objetivo era apresentar um rascunho preliminar da maneira pela qual a VEC poderia ser definida, além de alguns motivos segundo os quais uma abordagem nesses moldes seria necessária.

Muitas das ideias por trás da proposta da VEC têm análogos e antecedentes na literatura filosófica. Por exemplo, no campo da ética, as *teorias do observador ideal* procuram analisar conceitos normativos como "bom" ou "certo" em termos dos julgamentos que um observador ideal hipotético faria (onde um "observador ideal" se define como aquele que é onisciente sobre fatos não morais, é logicamente lúcido, é imparcial a respeito de questões relevantes, é livre de vários tipos de vieses e assim por diante).[454] A abordagem VEC, no entanto, não é (ou não precisa ser interpretada como) uma teoria moral. Ela não tem um compromisso com a afirmação de que há qualquer ligação entre valor e as preferências de nossas vontades extrapoladas coerentes. A VEC pode ser pensada simplesmente como uma maneira útil de aproximar qualquer coisa que tenha valor fundamental, ou pode ser considerada para além de qualquer conexão com a ética. Como ela é o principal protótipo da abordagem de normatividade indireta, vale a pena examiná-la mais detalhadamente.

Algumas explicações

Alguns termos encontrados na citação acima necessitam de uma explicação. "Pensássemos mais rápido", na terminologia de Yudkowsky, significa *se nós fôssemos mais espertos e tivéssemos pensado mais detalhadamente.* "Tivéssemos nos desenvolvido mais conjuntamente" parece significar *se nós tivéssemos aprendido, implementado melhorias em nossa cognição e nos autoaperfeiçoado sob condições adequadas de interação social uns com os outros.*

"Onde a extrapolação convergisse em vez de divergir" pode ser entendido da seguinte maneira: a IA deveria agir a respeito de alguma característica do resultado de sua extrapolação apenas se essa característica pudesse ser prevista pela IA com um grau razoavelmente alto de confiança. Uma vez que a IA não conseguiria prever o que nós desejaríamos caso fôssemos idealizados da maneira indicada, ela não deveria agir com base em uma suposição sem embasamento, mas, ao contrário, deveria se abster de agir. Entretanto, mesmo que muitos detalhes dos nossos desejos idealizados pudessem ser indeterminados ou imprevisíveis, poderiam haver alguns esboços desses desejos que a IA fosse capaz de compreender, sendo, desse modo, ao menos capaz de agir para garantir que o curso dos eventos futuros se desenrolasse de acordo com esses esboços. Por exemplo, se uma IA pudesse estimar confiavelmente que a nossa vontade extrapolada desejaria que nós não estivéssemos em agonia constante, ou que o universo não fosse coberto de clipes de papel, então a IA deveria agir para prevenir esses resultados.[455]

“Onde nossos desejos fossem coerentes em vez de interferir uns nos outros” pode ser lido da seguinte maneira: a IA deveria agir onde há um consenso razoavelmente amplo entre as vontades extrapoladas de humanos individuais. Um conjunto menor de desejos fortes e claros poderia, por vezes, superar os desejos fracos e confusos de uma maioria. Yudkowsky acredita que seria necessário um menor consenso para que a IA *evitasse* um resultado em particular que fosse estritamente especificado, e um maior consenso para que a IA agisse na condução do futuro em direção a alguma concepção específica e restrita de bondade. “A dinâmica inicial para a VEC”, ele descreve, “deveria ser conservadora para dizer ‘sim’ e prestar cuidadosa atenção aos ‘nãos’.”[456]

“Extrapolados da forma como gostaríamos que eles fossem extrapolados, interpretados da forma como gostaríamos que eles fossem interpretados”: a ideia por trás dessa última sentença parece ser a de que as regras para a extrapolação deveriam ser, elas mesmas, sensíveis à vontade extrapolada. Um indivíduo pode ter um desejo secundário (um desejo que diz respeito ao que desejar) de que alguns de seus desejos de primeira ordem não sejam levados em consideração quando sua vontade for extrapolada. Por exemplo, um alcoólatra que tenha um desejo de primeira ordem por bebidas alcoólicas pode também possuir o desejo secundário de não ter esse desejo de primeira ordem. Da mesma forma, nós podemos ter desejos a respeito de como as outras partes do processo de extrapolação deveriam se desenvolver, e esses desejos deveriam ser levados em consideração pelo processo de extrapolação.

Podemos contestar que, embora o conceito de vontade extrapolada coerente da humanidade pudesse ser apropriadamente definido, ainda assim seria impossível — até mesmo para uma superinteligência — descobrir o que a humanidade realmente desejaria sob as circunstâncias idealizadas hipotéticas estipuladas pela abordagem VEC. Sem nenhuma informação sobre o conteúdo de nossa vontade extrapolada, a IA estaria desprovida de qualquer padrão substancial para guiar seu comportamento. Todavia, apesar de ser difícil saber precisamente o que a VEC da humanidade desejaria, podemos fazer algumas previsões fundamentadas. Isso é possível inclusive hoje, na ausência de uma superinteligência. Por exemplo, é mais plausível que nossa VEC desejasse que houvesse pessoas no futuro que vivessem vidas enriquecedoras e felizes, enquanto é menos provável que ela desejasse que estivéssemos todos sentados em banquetas dentro de um quarto escuro sentindo dor. Se *nós* conseguimos ao menos efetuar alguns julgamentos de maneira sensata, uma superinteligência também o conseguiria. Desse modo, desde seus primeiros momentos, a conduta de uma superinteligência poderia ser guiada por suas estimativas a respeito do conteúdo da nossa VEC. Ela teria razões instrumentais fortes para refinar essas estimativas iniciais (por exemplo, estudando a cultura e a psicologia humanas, digitalizando cérebros humanos e refletindo sobre como poderíamos nos comportar caso tivéssemos mais conhecimento, pensássemos mais claramente etc.). Ao investigar esses assuntos, a IA seria guiada por suas estimativas iniciais em relação à nossa VEC; de modo que a IA, por exemplo, não decidiria rodar desnecessariamente uma infinidade de simulações

repletas de sofrimento humano caso ela estimasse que nossa VEC provavelmente reprovaria tais simulações por se tratarem de crimes da mente.

Outra objeção é que existe uma grande variedade de estilos de vida e códigos morais diferentes no mundo, sendo impraticável "misturá-los" em uma única VEC. Mesmo que alguém conseguisse misturá-los, o resultado poderia não ser particularmente saboroso — é impossível preparar um prato saboroso com a mistura dos melhores sabores encontrados nos pratos favoritos de todas as pessoas.[457] Em resposta, poderíamos apontar que a abordagem VEC não requer que todos os estilos de vida, códigos morais ou valores pessoais sejam misturados em um único prato. A dinâmica da VEC é supostamente agir apenas quando os nossos desejos forem coerentes. No caso de questões que geram um desacordo irreconciliável e generalizado, mesmo depois que as várias condições idealizadas tenham sido impostas, a dinâmica deveria ser abster-se de determinar o resultado. Para continuar com a analogia da culinária, indivíduos ou culturas podem ter pratos favoritos diferentes, mas, ainda assim, concordariam em relação a quais alimentos não seriam tóxicos. Desse modo, a dinâmica da VEC poderia agir no sentido de prevenir intoxicações alimentares, permitindo, ao mesmo tempo, que os humanos seguissem com suas práticas culinárias sem sua orientação ou interferência.

Argumentos para a VEC

O artigo de Yudkowsky apresenta sete argumentos para a abordagem VEC. Três deles são, basicamente, formas diferentes de enfatizar que, embora o objetivo deva ser fazer algo humanitário e benéfico, seria muito difícil estabelecer um conjunto de regras explícitas que não tivessem interpretações e consequências indesejadas.[458] A abordagem VEC deveria ser robusta e se autocorrigir; ela é feita para capturar a *origem* de nossos valores, e não para depender de que sejamos capazes de enumerar corretamente e articular, de uma vez por todas, cada um de nossos valores essenciais.

Os quatro argumentos restantes vão além desse primeiro ponto básico (mas importante), especificando as aspirações a respeito das soluções candidatas ao problema de especificação de valor e sugerindo que a VEC atenda a tais aspirações.

> *"Encapsular o crescimento moral"*

Essa é a aspiração de que a solução deveria permitir a possibilidade de progresso moral. Conforme sugerido anteriormente, há razões para acreditarmos que as nossas crenças morais atuais apresentam falhas de muitos tipos e mesmo que estejam profundamente equivocadas. Se estipulássemos um código moral específico e inalterável para ser seguido por uma IA, nós estaríamos, na verdade, cristalizando nossas convicções morais atuais, incluindo seus erros, e destruindo qualquer esperança de crescimento moral. A abordagem VEC, ao contrário, permite a possibilidade de tal crescimento porque deixa a cargo da IA a tentativa de fazer aquilo que desejaríamos que ela fizesse se tivéssemos nos desenvolvido mais e sob condições favoráveis, e é possível que, caso nosso desenvolvimento tivesse se

dado dessa forma, nossas crenças morais e sensibilidades estariam livres de seus defeitos e limitações atuais.

"Evitar sequestrar o destino da humanidade"

Yudkowsky tinha em mente um cenário no qual um pequeno grupo de programadores cria uma IA embrionária que se desenvolve em uma superinteligência e obtém uma vantagem estratégica decisiva. Nesse cenário, os programadores originais têm em suas mãos a herança cósmica de toda a humanidade. Obviamente, essa é uma responsabilidade terrível para qualquer mortal carregar em seus ombros. Mesmo assim, os programadores não podem fugir completamente dessa responsabilidade, pois eles se encontram nesta situação: qualquer escolha que eles façam, incluindo abandonar o projeto, teria consequências históricas globais. Yudkowsky vê a VEC como uma maneira pela qual os programadores evitariam assumir para si mesmos o privilégio ou o fardo de determinar o futuro da humanidade. Ao estabelecer uma dinâmica que implemente a vontade extrapolada coerente da *humanidade* — ao contrário de suas próprias vontades, ou de suas próprias teorias morais favoritas —, eles efetivamente repartem sua influência sobre o futuro com toda a humanidade.

"Evitar criar um motivo para que os humanos contemporâneos briguem pela dinâmica inicial"

Repartir a influência sobre o futuro da humanidade não é apenas algo moralmente preferível ao cenário de uma equipe de programadores implementando sua própria visão favorita, mas é também uma maneira de reduzir o incentivo de uma disputa em torno de quem criará a primeira superinteligência. Na abordagem VEC, os programadores (ou seus patrocinadores) não exercem maior influência sobre o conteúdo do resultado do que qualquer outra pessoa — embora desempenhem um papel causal relevante na determinação da estrutura da extrapolação e na decisão de implementar a VEC da humanidade em vez de outra alternativa qualquer. Evitar conflito é importante não apenas em virtude do dano imediato que um conflito tende a causar, mas também porque conflitos dificultam a colaboração no difícil desafio que é desenvolver uma superinteligência de maneira segura e benéfica.

A VEC foi pensada para ser capaz de receber um suporte amplo. Isso não se dá apenas porque ela distribui influência de modo igualitário. Há também uma base mais profunda para o potencial pacificador da VEC, que é, especificamente, o fato de que essa abordagem cria, em muitos grupos diferentes, a esperança de que suas visões preferidas de futuro poderão prevalecer totalmente. Imagine um membro do Talibã afegão debatendo com uma mulher que participa da Associação Humanista Sueca. Os dois têm visões de mundo muito diferentes, e o que é uma utopia para um pode ser uma distopia para o outro. Eles podem nem ao menos estar muito dispostos a chegar em algum tipo de consenso, como permitir que as meninas sejam escolarizadas, embora somente até o nono ano, ou permitir que as meninas suecas fossem escolarizadas, mas não as afegãs. Todavia, tanto o talibã quanto a

humanista poderiam ser capazes de endossar o princípio de que o futuro deveria ser determinado pela VEC da humanidade. O talibã poderia argumentar que caso seus pontos de vista religiosos estejam de fato corretos (conforme ele já se convenceu) e se há boas razões para aceitar que essas opiniões existam (como ele igualmente está convencido), então a humanidade, no fim das contas, aceitaria tais opiniões, caso as pessoas fossem menos preconceituosas e tendenciosas, se elas passassem mais tempo estudando as escrituras, se elas pudessem entender mais claramente como o mundo funciona e reconhecer suas prioridades essenciais, se elas pudessem se libertar de rebeldias e covardias irracionais e assim por diante.[459] A humanista, do mesmo modo, acreditaria que, sob essas mesmas condições idealizadas, a humanidade poderia abraçar os princípios adotados por ela.

"Manter a humanidade em última instância no controle do seu próprio destino"

Nós poderíamos não desejar um resultado no qual uma superinteligência paternalista ficasse nos observando constantemente e microgerenciando nossos afazeres na tentativa de otimizar cada detalhe de acordo com um grande plano. Mesmo que estipulássemos que a superinteligência fosse perfeitamente benevolente e livre de presunção, arrogância, prepotência, estreiteza de visão e outros defeitos humanos, ainda assim poderíamos nos ressentir da perda de autonomia vinculada a tal arranjo. Poderíamos preferir criar nosso próprio destino conforme seguíssemos adiante, mesmo que isso significasse nos atrapalhar pelo caminho. Talvez quiséssemos que a superinteligência servisse como uma rede de segurança para fornecer apoio quando as coisas saíssem catastroficamente erradas, mas, caso contrário, que nos deixasse cuidar de nossas próprias vidas.

A VEC permite essa possibilidade. Ela é pensada para ser uma "dinâmica inicial", um processo que se executa apenas uma vez e que depois substitui a si mesmo por qualquer que seja o desejo da vontade extrapolada. Se a vontade extrapolada da humanidade desejar que vivamos sob a supervisão paternalista de uma IA, então a dinâmica da VEC criaria essa IA e lhe passaria as rédeas. Se a vontade extrapolada, ao contrário, desejar que um governo democrático humano mundial seja criado, então a dinâmica da VEC poderia facilitar o estabelecimento de tal instituição e depois disso permanecer invisível. Se, diferentemente, a vontade extrapolada da humanidade seja a de que cada pessoa deve receber recursos que podem ser utilizados conforme preferir, desde que ela respeite os direitos iguais dos outros, então a dinâmica da VEC poderia transformar esse desejo em realidade operando nos bastidores como se fosse uma lei natural para prevenir invasões, roubo, assalto e outras interferências não consensuais.[460]

Desse modo, a estrutura da abordagem VEC permite uma gama virtualmente ilimitada de resultados. Também é concebível que a vontade extrapolada da humanidade desejasse que a VEC não fizesse absolutamente nada. Nesse caso, a IA responsável pela implementação da VEC deveria, após ter estabelecido com probabilidade suficiente que seria isso o que a vontade extrapolada da humanidade desejaria que ela fizesse, se desligar de maneira segura.

Observações adicionais

É claro que a proposta da VEC, como foi apresentada acima, é meramente esquemática. Ela possui uma série de parâmetros livres que poderiam ser especificados de várias formas, produzindo versões diferentes da proposta.

Um desses parâmetros é a base de extrapolação: de quem partiriam as vontades que seriam incluídas? Podemos dizer que seriam as vontades "de todos", mas essa resposta gera uma série de perguntas adicionais. A base de extrapolação inclui as chamadas "pessoas marginais", como os embriões, fetos, pessoas com morte cerebral, pacientes com demências severas, ou aqueles em estado vegetativo permanente? Cada hemisfério do cérebro de um paciente com a síndrome do cérebro dividido é levado em consideração na extrapolação e o peso que ele teria seria o mesmo atribuído ao cérebro inteiro de uma pessoa normal? E aquelas pessoas que viveram no passado, mas que agora estão mortas? Pessoas que nascerão no futuro? Animais mais evoluídos e outras criaturas conscientes? Mentes digitais? Extraterrestres?

Uma opção seria incluir apenas a população de todos os seres humanos adultos da Terra que estivessem vivos no momento da criação da IA. Uma extrapolação inicial partindo dessa base poderia, então, decidir se a base deveria ser expandida e de que maneira. Uma vez que os números de "marginais" na periferia dessa base é relativamente pequeno, o resultado da extrapolação pode não depender muito de onde exatamente essa fronteira seria demarcada — se, por exemplo, ela incluiria fetos ou não.

O fato de alguém ter sido excluído da base de extrapolação original não implica que seus desejos e bem-estar serão desconsiderados. Se a vontade extrapolada coerente daqueles que estão na base de extrapolação (por exemplo, seres humanos adultos vivos) desejar que considerações morais sejam estendidas para outros seres, então o resultado da dinâmica da VEC refletiria essa preferência. No entanto, é possível que os interesses daqueles que fossem incluídos na base da extrapolação original sejam melhor acomodados do que os interesses das pessoas de fora. Particularmente, caso a dinâmica venha a agir apenas se existir um acordo amplo entre vontades extrapoladas individuais (como na proposta original de Yudkowsky), parece haver um risco significativo de veto pouco generoso, que poderia impedir, por exemplo, que o bem-estar dos animais não humanos ou das mentes digitais fosse protegido. O resultado em potencial poderia ser moralmente detestável.[461]

Uma das motivações para a proposta da VEC foi impedir a criação de um pretexto para que os humanos disputassem entre si a criação da primeira IA superinteligente. Apesar dessa proposta se sair melhor nesse desiderato do que várias outras alternativas, ela não elimina por completo os motivos de conflito. Um indivíduo, grupo ou nação egoístas poderiam procurar aumentar sua fatia de ganhos no futuro mantendo outros fora da base de extrapolação.

Uma tomada de poder nesses moldes poderia ser racionalizada de várias maneiras. Podemos argumentar, por exemplo, que o patrocinador que financiasse o desenvolvimento da IA mereceria ser o dono de seus resultados. Enquanto afirmação moral, isso é provavelmente falso. Poderíamos contestar, por exemplo,

argumentando que o projeto que lançasse a primeira IA embrionária bem-sucedida estaria impondo uma imensa exposição a riscos ao restante da humanidade, a qual, portanto, mereceria ser compensada. A quantidade de compensação devida seria tão grande que todos deveriam ter participação nos resultados caso tudo corresse bem.[462]

Outro argumento que pode ser usado para racionalizar uma tomada de poder é o fato de que grandes segmentos da humanidade possuem preferências básicas ou malignas e que incluí-los na base de extrapolação arriscaria tornar o futuro da humanidade uma distopia. É difícil estabelecer as proporções de bondade e maldade presentes no coração de uma pessoa normal. É difícil também saber o quanto esse equilíbrio varia em diferentes grupos, camadas sociais, culturas ou nações. Mesmo que sejamos otimistas ou pessimistas a respeito da natureza humana, poderíamos preferir não apostar a herança cósmica da humanidade na especulação de que, entre a grande maioria dos 7 bilhões de pessoas atualmente vivas, prevaleceria o lado angelical em suas vontades extrapoladas. É claro que omitir um determinado grupo de pessoas da base da extrapolação não garante que a bondade triunfará e poderia ser que as almas que primeiro excluíssem outras ou tomassem o poder para si tendessem justamente a conter uma quantidade excepcionalmente alta de maldade.

Outro motivo de disputa pela dinâmica inicial se daria pelo fato de que alguém poderia acreditar que a IA de outra pessoa não funcionaria conforme o previsto, mesmo que essa IA tenha sido criada de modo a implementar a VEC na humanidade. Se diferentes grupos possuem crenças diferentes a respeito de qual implementação apresenta maiores chances de sucesso, poderia haver uma disputa para impedir que os demais lançassem a primeira IA. Em tal situação, se os projetos competidores pudessem resolver suas diferenças epistêmicas por meio de algum método que determinasse de maneira mais confiável quem está correto, seria uma solução muito melhor do que um conflito armado.[463]

Modelos de moralidade

A proposta da VEC não é a única forma possível de normatividade indireta. Por exemplo, em vez de implementar a vontade extrapolada coerente da humanidade, poderíamos tentar construir uma IA que tivesse como objetivo fazer o que é moralmente correto, confiando nas capacidades cognitivas superiores da IA para descobrir exatamente quais ações se encaixariam nessa descrição. Podemos chamar essa proposta de “retidão moral” (RM). A ideia é a de que nós, humanos, temos um entendimento imperfeito do que seja certo ou errado e talvez um entendimento ainda pior a respeito de como o conceito de retidão moral deva ser filosoficamente analisado, mas uma superinteligência poderia entender tudo isso de maneira mais apropriada.[464]

E se não tivermos certeza de que o realismo moral é verdadeiro? Ainda assim poderíamos tentar utilizar a proposta da RM. Para tanto, deveríamos apenas nos certificar de deixar especificado para a IA o que deveria ser feito na eventualidade de que sua pressuposição a respeito do realismo moral seja falsa. Por exemplo,

poderíamos estipular que se a IA estimasse com suficiente probabilidade que não existem verdades não relativas sobre retidão moral, então ela deveria passar a implementar a vontade extrapolada coerente ou simplesmente se desligar.[465]

A RM parece ter várias vantagens em relação à VEC. A RM eliminaria vários dos parâmetros livres da VEC, tais como o grau de coerência necessário entre vontades extrapoladas para que a IA agisse, a facilidade com a qual uma maioria poderia subjugar minorias discordantes e a natureza do ambiente social em que nossos egos extrapolados deveriam supostamente ter "evoluído em conjunto". Tudo indica que ela também eliminaria a possibilidade de uma falha moral resultante do uso de uma base de extrapolação que fosse muito limitada ou ampla demais. Além disso, a RM orientaria a IA em direção a uma ação moralmente correta mesmo se nossa vontade extrapolada coerente viesse a desejar que a IA tomasse medidas moralmente hediondas. Conforme mencionado, essa parece ser uma possibilidade bastante concreta na proposta da VEC. Bondade moral pode se assemelhar mais a um metal precioso do que a um elemento abundante na natureza humana, e mesmo depois que o minério tenha sido processado e refinado de acordo com as prescrições da proposta da VEC, quem pode garantir se o resultado principal será uma virtude resplandecente, uma escória indiferente ou um lixo tóxico?

A RM também parece ter algumas desvantagens. Ela se baseia na noção de algo "moralmente correto", um conceito notoriamente difícil, com o qual os filósofos têm se debatido desde a antiguidade sem nunca ter chegado a um consenso em suas análises. Escolher uma explicação equivocada para "retidão moral" poderia levar a resultados que seriam muito errados moralmente. Essa dificuldade na definição de "retidão moral" pode pesar muito contra o uso da proposta da RM. Entretanto, não está claro que essa proposta esteja realmente em desvantagem material nesse aspecto. A proposta da VEC também usa termos e conceitos que são difíceis de explicar (tais como "conhecimento", "mais parecidos com as pessoas que gostaríamos de ser", "nos desenvolvido mais conjuntamente", entre outros).[466] Mesmo se esses conceitos forem marginalmente menos opacos do que o de "retidão moral", eles estão ainda muito longe de qualquer coisa que os programadores possam atualmente expressar em forma de código.[467] O caminho para dotar uma IA com qualquer um desses conceitos pode envolver fornecer a ela habilidades linguísticas gerais (comparáveis, ao menos, àquelas de um humano adulto normal). Tal habilidade geral de compreensão da linguagem natural poderia, então, ser usada para entender o que significa "moralmente correto". Se a IA pudesse compreender seu significado, ela seria capaz de procurar por ações que se encaixassem nele. Na medida em que a IA desenvolvesse a superinteligência, ela poderia progredir em duas frentes: no problema filosófico de entender o que é a retidão moral e no problema prático de aplicar esse entendimento para avaliar ações específicas.[468] Embora não seja algo fácil, não é claro se seria *mais* difícil do que extrapolar a vontade extrapolada coerente da humanidade.[469]

Uma questão mais fundamental ligada à RM é que, ainda que possa ser implementada, essa abordagem poderia não nos oferecer aquilo que desejaríamos ou escolheríamos caso fôssemos mais inteligentes e melhor informados. Essa,

claro, é uma das características essenciais da RM, e não uma falha acidental. Contudo, ela poderia ser uma característica extremamente danosa para nós.[470]

Podemos tentar preservar a ideia básica do modelo RM ao mesmo tempo que reduzimos sua rigidez, focando na *permissibilidade moral*: poderíamos permitir que a IA seguisse a VEC da humanidade desde que ela não agisse de maneiras moralmente inadmissíveis. Por exemplo, o seguinte objetivo poderia ser formulado para a IA:

> Entre as ações moralmente admissíveis para a IA, escolha uma que a VEC da humanidade preferiria. Entretanto, se alguma parte dessa instrução não tiver um significado bem especificado, ou se estivermos demasiadamente confusos a respeito de seu significado, ou se o realismo moral for falso, ou se agimos de modo moralmente inaceitável ao criar uma IA com esse objetivo, então submeta-se a um desligamento controlado.[471] Siga o significado pretendido por essa instrução.

Ainda seria possível se preocupar com a possibilidade de que esse modelo de permissibilidade moral (PM) apresente um grau desagradavelmente alto de respeito aos requisitos de moralidade. O tamanho do sacrifício que ele exigiria depende de qual teoria ética é verdadeira.[472] Se a ética for *satisfatória*, no sentido de aceitar como moralmente permissível qualquer ação que satisfaça algumas restrições morais básicas, então a PM poderia deixar um amplo espaço para que nossa vontade extrapolada coerente influenciasse as ações da IA. Todavia, se a ética *for maximizadora* — por exemplo, se as únicas ações moralmente aceitáveis forem aquelas com as melhores consequências morais —, então a PM pode deixar pouco ou nenhum espaço para que nossas próprias preferências moldem o resultado.

Para ilustrar essa preocupação, retornemos por um momento ao exemplo do consequencialismo hedonista. Suponha que essa teoria ética seja verdadeira e que a IA saiba disso. Para os nossos propósitos imediatos, podemos definir o consequencialismo hedonista como a afirmação de que uma ação é moralmente correta (e moralmente admissível) se, e somente se, entre todas as ações possíveis, nenhuma outra ação produza mais prazer com menos sofrimento. Uma IA que seguisse a PM poderia maximizar o excesso de prazer convertendo o universo acessível em hedônio, um processo que poderia envolver a produção de computrônio para então usá-lo em cálculos que instanciassem experiências prazerosas. Uma vez que a simulação de qualquer cérebro humano existente não seria a maneira mais eficiente de produzir prazer, uma consequência provável dessa atitude é a de que todos nós morreríamos.

Ao adotarmos qualquer uma das abordagens, RM ou PM, estaríamos, dessa forma, arriscando sacrificar nossas vidas por um bem maior. Esse seria um sacrifício ainda maior do que parece, pois o que arriscamos perder não é meramente a chance de viver uma vida humana normal, mas a oportunidade de aproveitar as vidas muito mais longas e mais enriquecedoras que uma superinteligência amigável poderia nos proporcionar.

O sacrifício parece ainda menos atrativo quando nos damos conta de que a superinteligência, sacrificando muito menos do nosso próprio bem-estar potencial, poderia trazer benefícios quase tão bons (em termos fracionais). Suponha que concordemos em permitir que *quase* todo o universo acessível seja convertido em hedônio — tudo exceto uma pequena parte, digamos a Via Láctea, a qual seria deixada de lado para acomodar nossas necessidades. Ainda assim, teríamos 100 bilhões de galáxias dedicadas à maximização do prazer. Porém, teríamos uma galáxia na qual criaríamos civilizações maravilhosas que poderiam durar bilhões de anos e onde os humanos e animais não humanos poderiam sobreviver, prosperar e ter a oportunidade de se transformar em espíritos pós-humanos beatíficos.[473]

Se alguém escolhesse essa última opção (como eu tenderia a fazer), isso implicaria que essa pessoa não possui uma preferência lexicalmente dominante incondicional por agir de forma moralmente permissível. Mas essa opção corresponde a estipular um grande peso à moralidade.

Mesmo a partir de um ponto de vista puramente moral, poderia ser melhor *defender* alguma proposta que não seja tão moralmente ambiciosa quanto a RM ou a PM. Se o que for moralmente melhor não tiver chances de ser implementado — talvez em virtude de exigências rígidas —, poderia ser moralmente preferível promover alguma outra proposta que se aproximasse da ideal e cujas chances de implementação pudessem ser significantemente aumentadas se nós a promovêssemos.[474]

Faça o que eu quero dizer

Poderíamos nos sentir inseguros em relação à escolha de optar pela VEC, RM, PM ou algo diferente. Também não seria possível delegar essa decisão de alto nível, deixando a cargo da IA mais esse trabalho cognitivo? Qual é o limite para a nossa preguiça possível?

Considere, por exemplo, o seguinte objetivo “baseado em motivos”:

> Faça o que nos teria fornecido mais motivos para pedir que a IA o fizesse.

Esse objetivo pode se reduzir à vontade extrapolada, à moralidade ou a algo diferente, mas tudo indica que nos pouparia o esforço e o risco de tentar descobrir por nós mesmos quais desses objetivos mais específicos nos forneceriam maiores motivos de escolha.

Alguns dos problemas referentes aos objetivos baseados na moralidade, entretanto, também são aplicáveis aqui. Primeiro, poderíamos ter receio de que esses objetivos baseados em motivos deixassem pouquíssimo espaço para os nossos próprios desejos. Alguns filósofos defendem que uma pessoa sempre tem mais motivos para fazer o que seria moralmente melhor que ela fizesse. Se esses filósofos estiverem certos, o objetivo baseado em motivos seria equivalente à RM — com o risco concomitante de que a superinteligência que implementasse tal dinâmica matasse todos os que estivessem ao seu alcance. Segundo, como ocorre em todas as propostas formuladas em linguagem técnica, há a possibilidade de se

compreender incorretamente o significado de nossas próprias afirmações. Vimos que, no caso dos objetivos baseados na moralidade, pedir para que a IA faça o que é certo pode conduzir a consequências inesperadas e indesejadas, de tal modo que, se elas tivessem sido antecipadas por nós, não teríamos implementado o objetivo em questão. O mesmo se aplica se pedíssemos para que a IA fizesse aquilo que nos forneceria mais motivos para tal.

E se tentássemos evitar essas dificuldades formulando um objetivo em uma linguagem enfaticamente não técnica — por exemplo, usando termos relacionados à "gentileza":[475]

> Execute a ação mais gentil; ou, se nenhuma ação for a mais gentil,
> então execute uma ação que, pelo menos, seja maravilhosamente gentil.

Como poderia haver alguma objeção à construção de uma IA que fosse *gentil*? Mas devemos nos perguntar o que essa expressão significa precisamente. O léxico lista vários significados de "gentil" que claramente não fazem sentido nesse contexto: não desejamos que a IA seja *cordial e educada*, nem *extremamente delicada ou caprichosa.* Se pudermos ter certeza de que a IA reconhecerá a interpretação desejada para "gentileza" e pudesse ser motivada a exercer a gentileza apenas nesse sentido, então esse objetivo, ao que parece, equivaleria a um comando para que a IA fizesse o que os programadores realmente queriam dizer.[476] Uma ordem de efeito similar foi incluída na formulação da VEC ("... fossem [...] interpretados da forma como gostaríamos que eles fossem interpretados") e nos critérios de permissibilidade moral, conforme apresentados anteriormente ("... siga o significado pretendido por essa instrução"). Ao fixarmos a cláusula "Faça o que eu quero dizer", podemos indicar que as outras palavras presentes na descrição do objetivo deveriam ser interpretadas de forma benevolente, e não literalmente. Mas dizer que a IA deveria ser "gentil" não agrega quase nada: o verdadeiro trabalho é feito pela instrução "Faça o que eu quero dizer". Se soubéssemos como programar "Faça o que eu quero dizer" de uma maneira geral e poderosa, poderíamos muito bem definir isso como o objetivo único da IA.

Como poderíamos implementar uma dinâmica do tipo "Faça o que eu quero dizer"? Ou seja, como poderíamos criar uma IA motivada para interpretar de forma benevolente nossos desejos e intenções não verbalizadas e agir de acordo com eles? Um passo inicial poderia ser tentar tornar mais evidente o que queremos dizer com "Faça o que eu quero dizer". Talvez ajudasse se pudéssemos explicar isso usando termos mais behavioristas, por exemplo, em termos de preferências evidenciadas em várias situações hipotéticas — tais como situações nas quais tivéssemos mais tempo para considerar as opções, fôssemos mais inteligentes, soubéssemos mais a respeito de fatos relevantes e várias outras condições fossem mais favoráveis para manifestarmos precisamente, em termos de escolhas concretas, o que queremos dizer quando afirmamos que nosso desejo é uma IA que seja amigável, benéfica, gentil…

Nesse ponto, é claro, completamos o círculo. Acabamos retornando à abordagem da normatividade indireta com a qual iniciamos o capítulo — a proposta da VEC, que,

em sua essência, exclui todo o conteúdo concreto da especificação de valor, deixando apenas um valor abstrato definido em termos puramente procedurais: fazer aquilo que teríamos desejado que a IA fizesse em circunstâncias devidamente idealizadas. Através de tal normatividade indireta, poderíamos ter esperança de repassar à IA muito do trabalho cognitivo que nós mesmos estaríamos tentando desempenhar caso tentássemos articular uma descrição mais concreta a respeito de quais valores a IA deveria buscar. Por tentar tirar vantagem total da superioridade epistêmica da IA, a VEC poderia, dessa forma, ser vista como uma aplicação do princípio da deferência epistêmica.

Lista de componentes

Até agora consideramos diferentes opções em relação ao conteúdo que deveria ser implantado no sistema de objetivos. Mas o comportamento de uma IA também será influenciado por outras escolhas de projeto. Particularmente, a teoria da decisão e a epistemologia que seriam empregadas por uma IA poderiam fazer uma diferença crítica. Também seria importante questionar se os planos da IA estariam ou não sujeitos à revisão humana antes de serem colocados em prática.

A tabela 13 resume essas escolhas de projeto. Um projeto que almeja construir uma superinteligência deve ser capaz de explicar quais escolhas foram feitas em relação a cada um desses componentes e por que essas escolhas foram feitas.[477]

Conteúdo de objetivo

Já discutimos como a normatividade indireta pode ser usada na especificação de valores que IA deve buscar. Discutimos algumas opções, tais como os modelos baseados na moralidade e na vontade extrapolada coerente. Cada uma dessas opções demanda escolhas adicionais que precisariam ser feitas. Por exemplo, a abordagem da VEC pode apresentar diversas formas, dependendo de quem está incluído na base de extrapolação, da estrutura da extrapolação e assim por diante. Outras formas de métodos de seleção de motivação podem exigir diferentes tipos de conteúdo do objetivo. Por exemplo, um oráculo poderia ser construído para atribuir um valor ao fornecimento de respostas corretas. Um oráculo construído com motivação de domesticidade poderia também ter um conteúdo de objetivo que desvalorizasse o uso excessivo de recursos na elaboração de suas respostas.

TABELA 13

Lista de componentes

Conteúdo de objetivo	Qual objetivo a IA deveria buscar? Como uma descrição desse objetivo deveria ser interpretada? O objetivo deveria incluir a concessão de recompensas especiais para quem tivesse contribuído para o sucesso do projeto?
Teoria da decisão	A IA deveria usar a teoria da decisão causal, teoria da decisão evidencial, teoria da decisão sem atualizações ou alguma outra teoria?

Epistemologia Qual deveria ser a função de probabilidade *a priori* da IA e quais outras suposições explícitas ou implícitas ela deveria fazer a respeito do mundo? Qual teoria antrópica ela deveria usar?

Ratificação Os planos da IA deveriam ser submetidos à revisão humana antes de serem colocados em prática? Se esse for o caso, que protocolo seria utilizado no processo de revisão?

Outra opção de projeto poderia ser a inclusão de provisões especiais no conteúdo de objetivo como forma de recompensar indivíduos que tenham contribuído para o sucesso do desenvolvimento de uma IA, fornecendo-lhes, por exemplo, recursos adicionais ou uma maior influência sobre o comportamento da IA. Podemos chamar qualquer uma dessas provisões de "pacote de incentivos". O pacote de incentivos pode ser visto como uma maneira de aumentar a probabilidade de que o projeto seja bem-sucedido, à custa de, até certa medida, comprometer o objetivo que o projeto se propôs a alcançar.

Por exemplo, se o objetivo do projeto é criar uma dinâmica que implemente a vontade extrapolada coerente da humanidade, então um esquema de pacote de incentivos poderia especificar que deveria ser atribuído um peso adicional às vontades de alguns indivíduos na extrapolação. Se tal projeto for bem-sucedido, o resultado não será necessariamente a implementação da vontade extrapolada coerente da humanidade. Em vez disso, uma espécie de aproximação desse objetivo poderia ser alcançada.[478]

Uma vez que o pacote de incentivos seria uma parte do conteúdo de objetivo que deveria ser interpretado e seguido por uma superinteligência, ela poderia tirar vantagem da normatividade indireta para especificar provisões sutis e complicadas que seriam difíceis de ser implementadas por um gerente humano. Por exemplo, em vez de recompensar programadores de acordo com alguma métrica rudimentar, mas facilmente acessível, como quantas horas eles trabalharam ou quantos erros eles corrigiram, o pacote de incentivos poderia especificar que os programadores "deveriam ser recompensados de acordo com o aumento, permitido pelas suas contribuições, de alguma probabilidade prévia razoável de que o projeto fosse finalizado com sucesso da maneira que os patrocinadores pretendiam". Além disso, não haveria razão para limitar o pacote de incentivos apenas a membros do projeto. Em vez disso, poderia ser especificado que *cada* pessoa deveria ser recompensada de acordo com seus merecimentos. Alocação de crédito é um problema difícil, mas poderíamos esperar que uma superinteligência fosse capaz de realizar um trabalho razoável de aproximação dos critérios especificados, explícita ou implicitamente, pelo pacote de incentivos.

É até mesmo concebível que a superinteligência possa encontrar alguma maneira de recompensar indivíduos que tenham morrido antes de sua criação.[479] O pacote de incentivos poderia ser estendido de modo a incorporar pelo menos alguns desses falecidos, incluindo, potencialmente, indivíduos que tivessem morrido antes da concepção do projeto, ou mesmo cuja morte antecedesse a primeira enunciação do conceito de pacote de incentivos. Embora a instituição de tal política retroativa não pudesse incentivar causalmente aqueles que já se encontram descansando em seus túmulos enquanto essas palavras estão sendo digitadas, talvez ela possa ser favorecida por razões morais — embora possamos argumentar que, se a justiça

fosse um objetivo, ela deveria ser incluída como parte da própria especificação alvo, e não ser apenas uma parte do pacote de incentivos que a cerca.

Não podemos analisar aqui todos os assuntos éticos e estratégicos relacionados ao pacote de incentivos. A posição de um projeto a respeito dessas questões, todavia, seria um aspecto importante do conceito fundamental aplicado na criação do projeto.

Teoria da decisão

Outra escolha de projeto importante se refere à teoria da decisão seria atribuída à IA em sua construção. Isso poderia afetar a maneira como a IA se comporta em certas situações estratégicas cruciais. E poderia determinar, por exemplo, se a IA estaria pronta para negociar com (ou ser extorquida por) civilizações superinteligentes cuja existência seja conjecturada por ela. As particularidades da teoria da decisão também poderiam ter importância em situações difíceis que envolvessem probabilidades finitas de recompensas infinitas ("apostas de Pascal") ou probabilidades extremamente pequenas de pagamentos finitos extremamente grandes ("assaltos de Pascal") ou em contextos nos quais a IA esteja encarando uma incerteza normativa fundamental ou em que existam múltiplas instanciações do mesmo programa de agente.[480]

As opções disponíveis incluem a teoria da decisão casual (em suas diversas variações) e a teoria da decisão evidencial, acompanhada de outras candidatas mais novas, tais como a "teoria da decisão atemporal" e a "teoria da decisão sem atualizações", as quais ainda se encontram em desenvolvimento.[481] É provável que seja difícil identificar e articular a teoria da decisão correta, e que haja alguma dificuldade em confiar que tenhamos feito a escolha certa. Embora as perspectivas de que sejamos capazes de especificar diretamente uma teoria da decisão da IA sejam melhores do que a possibilidade de especificar diretamente seus objetivos finais, ainda assim somos confrontados com um risco substancial de erro. Muitas das complicações que talvez pudessem invalidar as teorias da decisão mais conhecidas atualmente foram descobertas há pouco tempo, sugerindo que talvez existam problemas adicionais que ainda não vieram à luz. O resultado de implantar uma teoria da decisão defeituosa em uma IA pode ser desastroso, levando, possivelmente, a uma catástrofe existencial.

Diante dessas dificuldades, poderíamos considerar uma abordagem indireta para especificar a teoria da decisão que uma IA deveria utilizar. Exatamente como fazer isso ainda não está muito claro. Talvez pudéssemos desejar que a IA usasse "a teoria da decisão *D* que teríamos desejado que ela usasse caso tivéssemos pensado longamente e cuidadosamente sobre o assunto". No entanto, a IA precisaria ser capaz de tomar decisões antes de aprender o que *D* é. Dessa maneira, seria necessário estabelecer alguma teoria da decisão efetiva e provisória *D'* que pudesse orientar a busca da IA por *D*. Poderíamos tentar definir *D'* como um tipo de superposição das hipóteses atuais da IA a respeito de *D* (ponderadas por suas probabilidades), embora existam problemas técnicos não solucionados a respeito da maneira pela qual isso poderia ser feito de uma forma totalmente geral.[482] Também há motivos para termos receio de que a IA pudesse tomar decisões ruins e

irreversíveis (tal como se reescrever para, a partir desse momento, operar com base em alguma teoria da decisão defeituosa) durante sua fase de aprendizagem, antes que IA tivesse tido oportunidade de determinar qual teria da decisão em particular seria a correta. Para reduzir o risco de que a IA saia dos trilhos durante esse período de vulnerabilidade, poderíamos, em vez disso, tentar fornecer à IA embrionária alguma forma de *racionalidade restrita*: uma teoria da decisão deliberadamente simplificada, a qual, porém, esperaríamos que fosse confiável e que ignorasse firmemente considerações esotéricas, mesmo aquelas que imaginássemos ser efetivamente legítimas, e que fosse projetada para substituir a si mesma por uma teoria da decisão mais sofisticada (especificada indiretamente) quando determinadas condições fossem preenchidas.[483] É uma questão de pesquisa ainda em aberto se e como isso poderia ser colocado em prática.

Epistemologia

Um projeto também precisará fazer uma escolha fundamental de concepção na seleção da epistemologia da IA, especificando os princípios e critérios segundo os quais as hipóteses empíricas deverão ser avaliadas. Dentro de uma arquitetura bayesiana, podemos pensar na epistemologia como uma função de probabilidade *a priori* — a atribuição implícita de probabilidades da IA para mundos possíveis antes que qualquer evidência perceptual tenha sido levada em consideração. Em outras arquiteturas, a epistemologia talvez assuma uma forma distinta; mas, de qualquer forma, alguma regra de aprendizagem indutiva seria necessária se a IA tiver que fazer generalizações baseadas em observações anteriores e apontar previsões a respeito do futuro.[484] Entretanto, assim como ocorre com o conteúdo de objetivo e a teoria da decisão, há um risco de que a nossa especificação da epistemologia venha a falhar.

Podemos imaginar que existe um limite para a quantidade de danos associada a uma especificação incorreta da epistemologia. Se a epistemologia for *demasiadamente* disfuncional, então a IA não poderia ser muito inteligente e também não poderia apresentar o tipo de risco discutido neste livro. Mas a preocupação é a de que nós poderíamos especificar uma epistemologia que seja suficientemente sólida, de modo a tornar a IA instrumentalmente efetiva para a maioria das situações, mas, mesmo assim, a epistemologia poderia apresentar alguma falha que faria com que a IA se perdesse em alguma questão crucial. Essa IA talvez possa ser semelhante a uma pessoa perspicaz cuja visão de mundo é baseada em um dogma falso, mantido com absoluta convicção. Essa pessoa, consequentemente, "ataca moinhos" e dá tudo de si para perseguir objetivos fantásticos ou perigosos.

Alguns tipos de diferença sutil nas probabilidades *a priori* de uma IA poderiam ter uma enorme influência na maneira como ela se comportaria. Por exemplo, uma IA poderia receber uma função de probabilidade *a priori* que indique que o universo apresenta uma probabilidade zero de ser infinito. Independentemente da quantidade de evidências astronômicas em contrário adquiridas por ela, tal IA teimosamente rejeitaria qualquer teoria cosmológica que apontasse para um universo infinito e, portanto, poderia fazer escolhas erradas.[485] Ou a IA poderia receber uma função de probabilidade *a priori* determinando que a probabilidade de que o universo seja

Turing computável é zero (essa é, de fato, uma característica comum de muitas funções de probabilidade *a priori* discutidas na literatura, incluindo a que utiliza a complexidade de Kolmogorov, mencionada no capítulo 1), novamente apresentando consequências pouco compreendidas caso a premissa aqui subentendida — conhecida como "Tese de Church-Turing" — for falsa. Uma IA poderia receber também uma função de probabilidade *a priori* que firmasse algum tipo de compromisso metafísico forte, por exemplo, excluindo *a priori* a possibilidade de que qualquer forma de dualismo forte entre corpo e mente pudesse ser verdadeira ou a possibilidade de que existam fatos morais irredutíveis. Se qualquer um desses compromissos estiverem errados, a IA talvez procurasse entender seus objetivos finais de maneiras que nós consideraríamos como instanciações perversas. No entanto, não existe uma razão óbvia para que essa IA, apesar de estar fundamentalmente errada sobre uma questão importante, pudesse não ser instrumentalmente efetiva o suficiente para adquirir uma vantagem estratégica decisiva. (A antropia, o estudo de como fazer inferências a partir de informações indexicais na presença da observação de efeitos de seleção, é outra área na qual a escolha dos axiomas epistêmicos poderia se provar essencial.)[486]

Talvez tenhamos razões suficientes para duvidar de nossa habilidade em resolver todos os assuntos epistemológicos fundamentais a tempo de construir a primeira IA embrionária. Podemos, portanto, considerar o uso de uma abordagem indireta para especificar a epistemologia da IA. Isso levantaria muitas das mesmas questões em relação a uma abordagem indireta para especificar a teoria da decisão de uma IA. Todavia, no caso da epistemologia, podemos ter uma esperança maior de que ocorra uma convergência benigna, com qualquer classe ampla de epistemologias sendo capaz de fornecer bases adequadas para a criação de uma IA segura e efetiva, produzindo, ao final, resultados doxásticos similares. A razão para isso é que evidências e análises empíricas suficientemente abundantes tenderiam a eliminar quaisquer diferenças moderadas nas expectativas prévias.[487]

Uma meta adequada seria dotar a IA com princípios epistemológicos fundamentais que coincidam com aqueles que governam nosso próprio pensamento. Qualquer IA que divergisse desse ideal seria uma IA que estaria raciocinando incorretamente, caso aplicássemos nossos próprios padrões de forma consistente. Isso, é claro, se aplica apenas aos nossos princípios epistemológicos *fundamentais*. Princípios não fundamentais deveriam ser continuamente criados e revisados pela própria IA embrionária na medida em que ela desenvolvesse sua compreensão a respeito do mundo. O propósito de uma superinteligência não é satisfazer preconceitos humanos, mas eliminar nossa ignorância e insensatez.

Ratificação

O último item da nossa lista de escolhas de projeto é a *ratificação*. Os planos da IA deveriam ser submetidos à revisão humana antes de serem colocados em prática? Para um oráculo, essa questão está implicitamente respondida de forma afirmativa. O oráculo provê informações e os revisores humanos escolhem se devem utilizá-las e de que maneira. No entanto, em relação aos gênios, soberanos e IAs

ferramentas, a questão em torno de se usar ou não alguma forma de ratificação continua em aberto.

Para ilustrar como a ratificação poderia funcionar, considere uma IA pensada para funcionar como um soberana implementando a VEC da humanidade. Em vez de utilizar diretamente essa IA, imagine que primeiramente construíssemos uma IA oráculo com o único propósito de responder perguntas que a IA soberano fizesse. Como os capítulos iniciais mostraram, existem riscos associados à criação de um oráculo superinteligente (por exemplo, riscos de crimes da mente ou de profusão de infraestrutura). Mas, para os propósitos desse exemplo, suponhamos que a IA oráculo tenha sido implementada com sucesso, de maneira que esses riscos tivessem sido evitados.

Teríamos, então, uma IA oráculo que nos daria suas melhores previsões a respeito das consequências de rodar um programa cujo objetivo fosse implementar a VEC da humanidade. O oráculo poderia não ser capaz de prever detalhadamente o que aconteceria, mas suas previsões seriam provavelmente melhores do que as nossas. (Se fosse impossível, mesmo para uma superinteligência, fazer *qualquer* previsão a respeito desse programa, seríamos loucos se o utilizássemos.) O oráculo, então, refletiria por um período de tempo e logo apresentaria suas previsões. Para fazer com que a resposta fosse inteligível, o oráculo talvez oferecesse ao operador uma série de ferramentas com as quais ele poderia explorar várias características do resultado previsto. O oráculo poderia mostrar imagens de como o futuro se pareceria e fornecer estatísticas a respeito da quantidade de seres conscientes que existirão em épocas diferentes, acompanhadas da média, do pico e dos menores níveis de bem-estar. Ele poderia apresentar biografias íntimas de muitos indivíduos selecionados aleatoriamente (talvez pessoas imaginárias selecionadas por sua provável representatividade). Ele poderia destacar aspectos do futuro que o operador não tivesse imaginado perguntar, mas que seriam considerados pertinentes uma vez que tivessem sido apontados.

Ser capaz de prever resultados dessa maneira fornece vantagens óbvias. A previsão poderia revelar as consequências de um erro nas especificações de projeto ou no código-fonte de um soberano planejado. Se a bola de cristal mostrasse um futuro em ruínas, poderíamos descartar o código-fonte do soberano planejado e tentar algo diferente. Podemos dizer que existem razões fortes para nos familiarizarmos com as ramificações concretas de uma opção antes de nos comprometermos com ela, especialmente quando todo o futuro da raça humana está em jogo.

Talvez seja menos óbvio que a ratificação também apresenta desvantagens potencialmente significativas. A qualidade pacificadora da VEC talvez possa ser abalada caso facções contrárias a ela, em vez de se submeterem à arbitragem de uma sabedoria superior na expectativa confiante de que receberiam uma confirmação, pudessem antever qual seria o veredicto. Alguém que propusesse a abordagem baseada na moralidade talvez estivesse preocupado com a possibilidade de que a determinação do patrocinador viesse por terra caso todos os sacrifícios necessários para a obtenção do moralmente ótimo fossem relevados. E talvez tenhamos toda a razão de preferir um futuro que traga algumas surpresas,

alguma dissonância, alguma selvageria, algumas oportunidades de que possamos nos superar — um futuro cujo contorno não esteja tão claramente delineado de modo a apresentar preconcepções, mas que dê espaço para algum movimento surpreendente e crescimento não planejado. Seríamos menos propensos a ter uma visão tão expansiva se pudéssemos escolher cada detalhe do futuro criteriosamente, enviando de volta à prancheta de desenho qualquer rascunho que não satisfizesse totalmente as nossas vontades naquele momento.

A questão em torno da ratificação por parte do patrocinador seria, então, menos evidente do que poderia parecer a princípio. Contudo, no geral, seria prudente nos aproveitarmos da oportunidade de pré-visualização, se essa funcionalidade estiver disponível. Mas em vez de deixar que a revisora ajustasse cada aspecto do resultado, talvez pudéssemos dar a ela um simples poder de veto, que poderia ser exercido apenas algumas vezes antes que o projeto fosse abortado por completo.[488]

Chegando perto o suficiente

O propósito principal da ratificação seria reduzir a probabilidade de um erro catastrófico. De um modo geral, parece ser uma atitude sábia almejar a diminuição do risco de um erro catastrófico em vez de maximizar as chances de que cada detalhe seja totalmente otimizado. Há duas razões para isso. Primeiro, a herança cósmica da humanidade é astronomicamente grande — há recursos o suficiente, mesmo que esse processo envolva algum desperdício ou aceite algumas restrições desnecessárias. Segundo, há a esperança de que, caso sejamos capazes de alcançar as condições iniciais para uma explosão de inteligência de maneira aproximadamente correta, a superinteligência resultante talvez possa, finalmente, focar e alcançar precisamente nossos objetivos últimos. O importante é aterrissar na bacia certa de atratores.

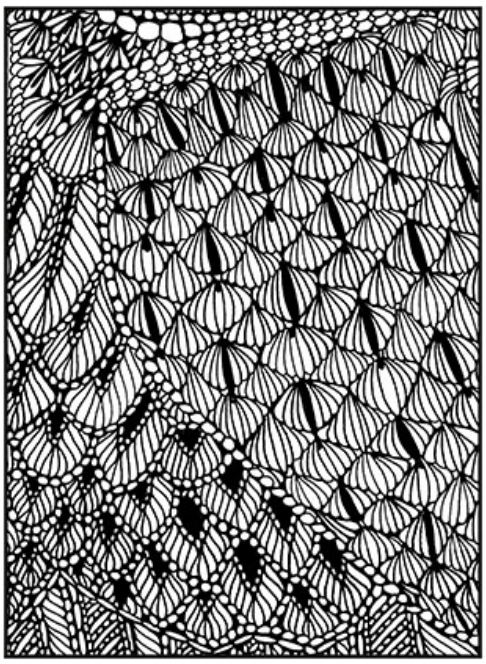
Em relação à epistemologia, é possível que uma ampla gama de funções de probabilidade *a priori* venha a convergir, em última instância, para posteriores similares (quando calculados por uma superinteligência e condicionados a uma quantidade realística de dados). Logo, não precisamos nos preocupar em obter a epistemologia de maneira *exatamente* correta. Apenas seria necessário evitar que sejam fornecidas à IA probabilidades *a priori* extremas a ponto de torná-la incapaz de aprender verdades vitais, mesmo que ela tenha uma larga experiência e uma grande capacidade de análise.[489]

Quanto à teoria da decisão, o risco de um erro irrecuperável parece ser maior. Talvez ainda possamos ter a esperança de conseguir especificar diretamente uma teoria da decisão que seja boa o suficiente. Uma IA superinteligente poderia mudar para uma nova teoria da decisão a qualquer momento; entretanto se essa IA iniciar com uma teoria da decisão suficientemente errada, ela pode não ver razão para essa troca. Mesmo que um agente perceba os benefícios de possuir uma teoria da decisão diferente, essa percepção pode ocorrer tarde demais. Por exemplo, um agente concebido para rejeitar chantagens pode se beneficiar caso consiga dissuadir potenciais chantagistas. Por essa razão, agentes passíveis de serem

chantageados poderiam fazer bem em proativamente adotar uma teoria da decisão que não fosse passível de explorações. No entanto, assim que um agente passível de ser chantageado recebesse uma ameaça e a considerasse crível, o estrago estaria feito.

Uma vez fornecidas uma epistemologia e uma teoria da decisão adequadas, poderíamos tentar projetar o sistema de maneira que ele implementasse a VEC ou algum outro conteúdo de objetivo indiretamente especificado. Novamente há uma esperança de convergência, ou seja, de que diferentes maneiras de implementar uma dinâmica semelhante à VEC conduziriam ao mesmo resultado utópico. Na ausência de tal convergência, nós ainda poderíamos esperar que muitos dos diferentes resultados possíveis seriam bons o suficiente para contar como um sucesso existencial.

Não é necessário criar um projeto altamente otimizado. Nosso foco, ao contrário, deveria se manter na criação de um projeto altamente confiável, o qual nos passaria a confiança de ter sanidade suficiente para reconhecer suas próprias falhas. Uma superinteligência imperfeita, cujos fundamentos sejam sólidos, gradualmente consertaria a si mesma e, assim, ela exerceria a mesma quantidade de poder benéfico de otimização no mundo se tivesse sido perfeita desde o início.



14

CAPÍTULO 14

O PANORAMA ESTRATÉGICO

Agora é o momento de considerarmos o desafio da superinteligência em um contexto mais amplo. Gostaríamos de ser capazes de nos orientar suficientemente no cenário estratégico para saber, pelo menos, qual direção geral devemos seguir. Isso, efetivamente, não é algo simples. Aqui, neste penúltimo capítulo, introduziremos alguns conceitos analíticos gerais que nos ajudarão a pensar a respeito de políticas científicas e tecnológicas de longo prazo. Depois, aplicaremos esses conceitos à questão da inteligência de máquina.

Seria esclarecedor fazer uma distinção aproximada entre duas posições normativas diferentes a partir das quais uma política proposta poderia ser avaliada. A *perspectiva do indivíduo-afetado* questiona se uma determinada mudança seria do "nosso interesse" — ou seja, se despertaria (no equilíbrio e em expectativa) o interesse de criaturas dotadas de status moral já existentes ou que virão a existir independentemente da ocorrência ou não dessa mudança proposta. A *perspectiva impessoal*, ao contrário, não confere nenhuma consideração especial aos indivíduos já existentes ou àqueles que virão a existir independentemente da ocorrência dessa mudança proposta. Em vez disso, considera todos igualmente, não obstante sua localização temporal. A perspectiva impessoal reconhece um grande valor em trazer novos indivíduos à existência, conquanto eles tenham vidas que valham a pena ser vividas: quanto mais vidas felizes forem criadas, melhor.

Essa distinção, apesar de apenas insinuar as complexidades morais associadas a uma revolução de inteligência de máquina, pode ser útil em uma análise preliminar. Aqui examinaremos primeiro as questões a partir de uma perspectiva impessoal. Feito isso, analisaremos o que poderia mudar caso uma ênfase fosse dada à perspectiva do indivíduo-afetado em nossas deliberações.

Estratégia científica e tecnológica

Antes de nos concentrarmos em questões específicas a uma superinteligência de máquina, é necessário que sejam introduzidos alguns conceitos e considerações estratégicos relacionados ao desenvolvimento científico e tecnológico de maneira mais geral.

Desenvolvimento tecnológico diferencial

Suponha que um político proponha que cortemos o financiamento de alguma área de pesquisa por conta dos riscos e consequências de longo prazo de alguma tecnologia hipotética que poderia ser produzida por meio do uso de conhecimento

dessa área. Esse político certamente poderia esperar uma grande oposição por parte da comunidade científica.

Os cientistas e seus defensores costumam dizer que é inútil impedir a realização de pesquisas como tentativa de conter a evolução tecnológica. Se uma tecnologia é viável (segundo esse argumento), ela será desenvolvida independentemente de qualquer escrúpulo que políticos possam ter em relação a riscos futuros especulativos. De fato, quanto mais poderosas forem as capacidades que esperamos produzir por meio de uma certa linha de pesquisa e desenvolvimento, mais certeza podemos ter de que alguém, em algum lugar, terá motivos para segui-la. Cortes nos financiamentos não deteriam o progresso ou evitariam seus perigos concomitantes.

O mais interessante é que essa objeção baseada no argumento da futilidade quase nunca é levantada quando um político propõe *aumentar* o financiamento em alguma área de pesquisa, embora esse argumento pareça válido para ambos os casos. Raramente ouvimos vozes indignadas protestando: “Por favor, não aumentem nosso financiamento. Em vez disso, façam alguns cortes. Pesquisadores em outros países certamente tomarão nosso lugar, e esse mesmo trabalho será feito de qualquer maneira. Não desperdicem o dinheiro do contribuinte em pesquisas científicas nacionais!”.

O que explica essa aparente contradição? Uma explicação plausível, obviamente, é que os membros da comunidade científica apresentam um viés de autoconveniência que nos faz acreditar que pesquisas são sempre benéficas, levando-nos a apoiar qualquer argumento que dê sustentação à nossa demanda por mais financiamento. No entanto, esse duplo padrão pode ser explicado em termos de proteção dos interesses nacionais. Suponhamos que o desenvolvimento de uma tecnologia tenha *dois* efeitos: conceder um pequeno benefício *B* a seus inventores e ao país que os financiou e impor um dano *H* agregadamente maior — que poderia ser um risco de externalidade — a todo o conjunto da população. Mesmo alguém que seja extremamente altruísta poderia optar pelo desenvolvimento dessa tecnologia globalmente danosa. Seria possível chegar à conclusão de que o risco *H* viria a ocorrer independentemente do que fosse feito, já que, caso um determinado grupo se abstivesse de desenvolver a tecnologia, algum outro poderia vir a fazê-lo; e considerando que o bem-estar geral não poderia ser afetado, o grupo poderia muito bem optar por obter o benefício *B* para seus integrantes e sua nação. (“Infelizmente, muito em breve existirá um dispositivo que destruirá o mundo. Felizmente, nós temos verbas para criá-lo!”)

Qualquer que seja a explicação para o apelo da objeção baseada no argumento da futilidade, não seria possível demonstrar que, em geral, inexiste uma razão impessoal nas tentativas de conduzir o desenvolvimento tecnológico. Há uma falha até mesmo quando reconhecemos a ideia motivadora de que, com esforços continuados de desenvolvimento científico e tecnológico, todas as tecnologias relevantes seriam, eventualmente, desenvolvidas — ou seja, mesmo que concordemos com as afirmações que virão a seguir.

Conjectura de realização tecnológica

> Se os esforços para o desenvolvimento científico e tecnológico não forem encerrados efetivamente, então todas as capacidades básicas importantes que poderiam ser obtidas por meio de alguma tecnologia possível o serão.[490]

Existem ao menos duas razões pelas quais a conjectura de realização tecnológica não implica na objeção baseada no argumento da futilidade. Primeiro, o antecedente pode não ser verdadeiro porque não é possível saber se os esforços para o desenvolvimento científico e tecnológico não se encerrarão efetivamente (antes que fôssemos capazes de alcançar a maturidade tecnológica). Essa ressalva é especialmente pertinente dentro de um contexto que envolva riscos existenciais. Segundo, mesmo que tivéssemos certeza de que viriam a ser obtidas todas as capacidades básicas importantes passíveis de se obter por meio de alguma tecnologia possível, ainda assim poderia fazer sentido uma tentativa de influenciar o direcionamento das pesquisas científicas. O que importa não é apenas *se* uma tecnologia será desenvolvida, mas também *quando* isso ocorrerá, *quem* a desenvolverá e *em que contexto*. As circunstâncias do surgimento de uma nova tecnologia, as quais influenciam seus impactos, poderiam ser alteradas por meio de aumentos e reduções nos fluxos de financiamento das pesquisas (e através de outras políticas de incentivo).

Essas reflexões sugerem um princípio que nos levaria a atentar para a velocidade relativa com a qual cada tecnologia diferente seria desenvolvida:[491]

> **O princípio do desenvolvimento tecnológico diferencial**
> Atrase o desenvolvimento de tecnologias perigosas e prejudicais, em especial as que aumentem o nível de risco existencial, e acelere o desenvolvimento de tecnologias benéficas, especialmente as que reduzem os riscos existenciais impostos pela natureza ou por outras tecnologias.

Desse modo, uma política poderia ser avaliada com base na quantidade de vantagens diferenciais que ela concede a formas desejáveis de desenvolvimento tecnológico, em comparação às formas indesejáveis.[492]

Ordem preferível de chegada

Algumas tecnologias têm um efeito ambivalente em relação a riscos existenciais, aumentando alguns desses riscos conforme diminui outros. A superinteligência é uma delas.

Vimos, nos capítulos anteriores, que a introdução da superinteligência de máquina poderia acarretar um risco existencial substancial. Porém, também reduziria diversos outros riscos existenciais. Riscos relacionados a causas naturais — como impacto de asteroides, supervulcões e pandemias naturais — seriam virtualmente eliminados, uma vez que a superinteligência poderia empregar contramedidas para combater a grande maioria desses perigos, ou, pelo menos, demovê-los à categoria de não existenciais (através da colonização do espaço, por exemplo).

Esses riscos existenciais naturais são comparativamente pequenos quando vistos a partir de uma escala relevante de tempo. Mas uma superinteligência também eliminaria ou reduziria muitos riscos de origem antropogênica. Ela reduziria, particularmente, os riscos de destruição acidental, incluindo aqui riscos de acidentes relacionados a novas tecnologias. Como, em geral, ela seria mais capacitada que os humanos, uma superinteligência provavelmente teria menos chances de cometer erros, estaria mais propensa a reconhecer quando as precauções seriam necessárias e seria capaz de implementar essas precauções de modo competente. Uma superinteligência bem construída poderia, por vezes, decidir correr riscos, mas apenas se essa fosse uma atitude sensata. Além do mais, ao menos em cenários nos quais a superinteligência forme um *singleton*, muitos riscos existenciais antropogênicos não acidentais, derivados de problemas de coordenação global, seriam eliminados. Entre eles, os riscos de guerras, corridas tecnológicas, formas indesejáveis de competição e evolução, bem com as "tragédias dos comuns".

Uma vez que perigos substanciais estão associados ao desenvolvimento, pelos seres humanos, de biologia sintética, nanotecnologia molecular, engenharia climática, instrumentos para melhorias biomédicas, manipulação neuropsicológica, ferramentas para controle social que pudessem facilitar o surgimento de regimes tiranos ou totalitários e outras tecnologias ainda não imaginadas, a eliminação de todos esses tipos de riscos seria uma grande dádiva. Poderíamos, desse modo, argumentar que quanto mais cedo se der a criação de uma superinteligência, melhor. Entretanto, se riscos advindos da natureza ou outros riscos não relacionados a tecnologias futuras são pequenos, então esse argumento poderia ser refinado: o que importa é que alcancemos a superinteligência *antes* que conquistemos outras tecnologias perigosas, como a nanotecnologia avançada. Pode não importar se isso acontecer mais cedo ou mais tarde (com base em uma perspectiva impessoal), contanto que a ordem esteja correta.

A afirmação de que é preferível que uma superinteligência surja antes de outras tecnologias potencialmente perigosas, como a nanotecnologia, está baseada no fato de que uma superinteligência reduziria os riscos existenciais ligados à nanotecnologia, mas o contrário não ocorreria.[493] Consequentemente, se criarmos primeiro uma superinteligência, enfrentaremos apenas os riscos existenciais associados a uma superinteligência; por outro lado, se a nanotecnologia for criada primeiro, enfrentaremos os riscos associados à nanotecnologia e, depois disso, os riscos associados à superinteligência.[494] Mesmo que os riscos existenciais relacionados a uma superinteligência sejam muito grandes e que a superinteligência seja a mais perigosa de todas as tecnologias, ainda assim poderia haver um motivo para que apressássemos o seu desenvolvimento.

No entanto, argumentos do tipo "quanto antes melhor" pressupõem que o grau de risco associado à criação de uma superinteligência é o mesmo, independentemente do momento em que ela for criada. Se, caso contrário, esse risco diminuísse com o passar do tempo, seria mais vantajoso atrasar o advento da revolução de inteligência de máquina. Mesmo que um surgimento mais tardio da superinteligência permitisse que outras catástrofes existenciais tivessem mais chance de ocorrer, ainda assim poderia ser preferível atrasar o desenvolvimento de

uma superinteligência. Isso seria especialmente plausível se os riscos existenciais associados a uma superinteligência forem muito maiores do que aqueles associados a outras tecnologias disruptivas.

Existem muitas razões fortes para acreditarmos que o grau de risco associado a uma explosão de inteligência diminuiria significativamente ao longo de um período de muitas décadas. Umas delas é que uma data mais à frente nos daria mais tempo para desenvolver soluções em relação ao problema do controle. O problema do controle foi reconhecido apenas recentemente e muitas das melhores ideias a respeito de como abordá-lo foram descobertas apenas em torno da última década (e, em muitos casos, durante o período em que este livro estava sendo escrito). É provável que o estado da arte avance muito ao longo das próximas décadas, e caso o problema se mostre muito difícil de ser solucionado, um nível significativo de progresso pode continuar a ocorrer por um século ou mais. Quanto mais demorado for o surgimento de uma superinteligência, maior será o progresso que terá sido alcançado quando isso acontecer. Essa é uma consideração importante em favor do seu surgimento mais futuramente — e uma consideração muito forte contra o seu surgimento em datas muito próximas ao presente.

O surgimento tardio de uma superinteligência também poderia ser mais seguro porque haveria mais tempo para que um quadro geral de avanços benéficos pudesse ser alcançado pela civilização humana. O peso que damos a essa consideração dependerá do quão otimistas somos em relação a tais tendências gerais.

Uma pessoa otimista poderia certamente apontar para um grande número de indicadores animadores e diversas possibilidades esperançosas. As pessoas poderiam ter aprendido a se relacionar melhor umas com as outras, o que reduziria a violência, as guerras e a crueldade; a coordenação global e o nível de integração política poderiam aumentar, o que tornaria mais fácil evitar corridas tecnológicas indesejáveis (mais sobre isso adiante) ou ainda permitiria a negociação de acordos nos quais os possíveis ganhos derivados de uma explosão de inteligência pudessem vir a ser divididos igualmente. Aparentemente, há uma tendência histórica de longo prazo que aponta nessa direção.[495]

Além disso, um otimista pode acreditar que o “nível de sanidade” da humanidade aumentará com o passar dos séculos — que os preconceitos (como um todo) diminuirão, que muitas boas ideias terão sido acumuladas e que as pessoas se tornarão mais habituadas a pensar em probabilidades abstratas em relação a acontecimentos futuros e riscos globais. Com sorte, poderíamos ver um aumento geral nos padrões epistêmicos, tanto na cognição individual como na coletiva. Novamente, existem tendências que apontam nessa direção. O progresso científico sugere um maior conhecimento a respeito de assuntos mais variados. O crescimento econômico pode oferecer nutrição adequada a uma proporção maior da população global (principalmente durante os primeiros anos de vida, que são cruciais para o desenvolvimento cerebral) e um acesso mais amplo à educação de qualidade. Avanços em tecnologia da informação tornarão mais fácil encontrar, integrar, avaliar e comunicar dados e ideias. Além disso, até o final deste século, a

humanidade terá passado mais cem anos cometendo erros, a partir dos quais algo poderá ser aprendido.

Muitos desenvolvimentos tecnológicos em potencial são ambivalentes quando considerados segundo os critérios acima mencionados — aumentando alguns riscos existenciais e diminuindo outros. Por exemplo, avanços em técnicas de vigilância, mineração de dados, detecção de mentiras, biometria e a possibilidade de manipulação de crenças e desejos por meios psicológicos e neuroquímicos poderiam reduzir alguns riscos existenciais através da simplificação da coordenação internacional ou da eliminação do terrorismo e de grupos rebeldes. No entanto, esses mesmos avanços também poderiam aumentar alguns riscos existenciais com a amplificação de dinâmicas sociais indesejáveis ou permitindo a formação de regimes totalitários permanentemente estáveis.

Outra fronteira importante seria a do aperfeiçoamento da cognição biológica, por exemplo, através da seleção genética. Quando discutimos esse assunto nos capítulos 2 e 3, chegamos à conclusão de que as formas mais radicais de superinteligência surgiriam, mais provavelmente, por meio de algum tipo de inteligência de máquina. Essa afirmação é coerente com um cenário em que o aperfeiçoamento biológico desempenhasse um papel importante no planejamento e criação de uma superinteligência de máquina. As melhorias genéticas, obviamente, parecem ser uma boa maneira de reduzir riscos: quanto mais inteligentes forem as pessoas trabalhando na resolução do problema do controle, maiores seriam as chances de se encontrar uma solução. Entretanto, o aperfeiçoamento cognitivo também poderia acelerar o desenvolvimento da inteligência de máquina, reduzindo, assim, o tempo disponível para a solução do problema do controle. O aperfeiçoamento cognitivo também teria muitas outras consequências relevantes. Essas questões merecem ser analisadas mais detalhadamente. (A maioria das afirmações a respeito do “aperfeiçoamento cognitivo” feitas a seguir se aplica igualmente a meios não biológicos através dos quais poderíamos aumentar nossa efetividade epistêmica coletiva e individual.)

Taxas de mudança e aperfeiçoamento cognitivo

Um aumento na média ou na faixa superior da habilidade intelectual humana provavelmente aceleraria o progresso tecnológico em todas as áreas, incluindo o progresso em direção a diversas formas de inteligência de máquina, o progresso em relação à resolução do problema do controle e o progresso em um gama variada de outros objetivos técnicos e econômicos. Qual seria o resultado líquido de tal aceleração?

Considere o caso limite de um “acelerador universal”, uma intervenção imaginária que, literalmente, aceleraria *tudo o que existe*. As ações de tal acelerador universal corresponderiam apenas a um redimensionamento arbitrário em nossa medida de tempo e não produziriam uma mudança qualitativa nos resultados observados.[496]

Se quisermos entender a ideia de que o aperfeiçoamento cognitivo poderia, de um modo geral, acelerar as coisas, nós certamente precisaremos de outro conceito que não o do acelerador universal. Uma abordagem mais promissora consiste em nos concentrar no quão capaz o aperfeiçoamento cognitivo seria em aumentar a

taxa de mudanças em um tipo de processo *relativamente* à taxa de mudanças em algum outro tipo de processo. Tal aceleração diferencial poderia afetar a dinâmica de um sistema. Desse modo, considere o seguinte conceito:

> *Acelerador de desenvolvimento macroestrutural* — uma alavanca que acelera a taxa na qual características macroestruturais da condição humana se desenvolvem, enquanto mantém imutáveis a taxa em que os acontecimentos humanos de nível micro se desenrolam.

Imagine que essa alavanca fosse colocada na posição de "desacelerar" e isso desencadeasse um mecanismo que freasse a grande roda da história, em meio a muitas faíscas e ruídos metálicos. Depois que a roda se acomodasse em um ritmo mais lento, o resultado seria um mundo onde a inovação tecnológica ocorreria num ritmo mais calmo e no qual mudanças fundamentais ou globalmente significativas na estrutura política e na cultura ocorreriam com menos frequência e de forma menos abrupta. Um número maior de gerações transcorreria antes que uma era desse lugar a outra. Uma pessoa, durante seu curso de vida, veria poucas mudanças na estrutura básica das condições humanas.

Durante a maior parte da existência de nossa espécie, o desenvolvimento macroestrutural se deu de forma mais lenta do que ocorre hoje em dia. Há 50 mil anos, um milênio poderia transcorrer por completo sem o surgimento de uma única invenção tecnológica significativa, sem qualquer aumento perceptível do conhecimento e entendimento humanos e sem qualquer mudança política global significativa. Todavia, em um nível micro, o caleidoscópio de acontecimentos humanos se movimentava num ritmo razoável, com nascimentos, mortes e outros eventos pessoal e localmente significativos. O dia de uma pessoa comum que viveu no Plistoceno talvez fosse muito mais agitado do que nossos dias são atualmente.

Se você encontrasse uma alavanca mágica que permitisse mudar a velocidade do desenvolvimento macroestrutural, o que você deveria fazer? Você deveria acelerá-lo, desacelerá-lo ou deixar as coisas como estão?

Assumindo o ponto de vista impessoal, essa questão requer que consideremos os efeitos dessa mudança em relação aos riscos existenciais. Façamos uma distinção entre dois tipos de riscos: "riscos de situação" e "riscos de etapa". Um risco de situação é aquele associado ao fato de nos encontrarmos em uma determinada situação, e a quantidade total de risco de situação a que o sistema estaria exposto é uma função direta do período no qual o sistema se mantém nessa situação. Riscos naturais são, tipicamente, riscos de situação: quanto mais tempo ficarmos expostos, maior será a chance de sermos atingidos por um asteroide, uma erupção de um supervulcão, uma tempestade radioativa, pandemias naturais ou algum outro golpe da foice cósmica. Alguns riscos antropogênicos também são riscos de situação. Em nível individual, quanto mais tempo um soldado ficar com a cabeça para fora da trincheira, maior será a chance cumulativa de ser atingido pelo inimigo. Também existem riscos de situação antropogênicos de nível existencial: quanto mais tempo vivermos em um sistema internacionalmente arcaico, maior será a chance cumulativa de um armagedom termonuclear ou de uma grande guerra

com outros tipos de armas de destruição em massa, o que acabaria com a civilização.

Um risco de etapa, por outro lado, é um risco discreto, associado a alguma transição necessária ou desejável. Uma vez que essa transição tenha se completado, esse risco desaparece. A quantidade de risco de etapa associado a uma determinada transição não é, em geral, uma função simples do tempo que essa transição leva para se completar. O risco de se atravessar um campo minado não diminui pela metade se o fizermos no dobro da velocidade. Caso ocorra por meio de uma partida rápida, a criação de uma superinteligência poderia representar um risco de etapa: existiria um certo risco associado à partida, sua magnitude dependeria das precauções que tivessem sido tomadas; mas a quantidade de risco poderia não depender muito da duração da partida: se esta ocorreu em vinte milissegundos ou vinte horas.

Desse modo, podemos afirmar o seguinte em relação a um hipotético acelerador de desenvolvimento macroestrutural:

- Levando em conta os riscos existenciais de situação, deveríamos ser favoráveis à aceleração — contanto que imaginemos a existência de uma perspectiva realista de que chegaremos a uma era pós-transição, na qual a presença de outros riscos existenciais seria muito menor.
- Se soubéssemos da existência de alguma etapa futura que estivesse destinada a causar uma catástrofe existencial, nesse caso, deveríamos desacelerar o ritmo de desenvolvimento macroestrutural (ou mesmo revertê-lo) para que mais gerações tenham tempo de existir antes que as cortinas caiam. Mas, em realidade, seria demasiadamente pessimista ter tanta certeza de que a humanidade estaria condenada.
- No momento, o nível de risco de situação existencial parece ser relativamente baixo. Se imaginarmos as macrocondições tecnológicas da humanidade congeladas em seu estágio atual, parece muito pouco provável que uma catástrofe existencial pudesse acontecer em uma escala de tempo de, digamos, uma década. Desse modo, um atraso de uma década — conquanto que ocorresse em nosso atual estágio de desenvolvimento, ou em algum outro momento que apresentasse um baixo risco de situação — incorreria em um risco de situação muito pequeno, enquanto o adiamento de uma década nos desenvolvimentos tecnológicos subsequentes poderia muito bem ter um impacto positivo significativo nos riscos existenciais de etapas posteriores, permitindo, por exemplo, mais tempo de preparação.

Ou seja: a velocidade de desenvolvimento macroestrutural é importante, principalmente, em virtude da sua influência no grau de preparação que a humanidade teria atingido quando chegasse o momento de enfrentar os principais riscos de etapa.[497]

Sendo assim, deveríamos nos questionar até que ponto o aperfeiçoamento cognitivo (e a concomitante aceleração do desenvolvimento macroestrutural) afetaria o nível esperado de preparação no momento crítico. Seria preferível um

período curto de preparação com uma inteligência maior? Com uma inteligência maior, o período de preparação poderia ser usado de forma mais efetiva e o passo crítico final seria dado por uma humanidade mais inteligente. Ou, se assim pudermos obter mais tempo de preparação, seria preferível operar com níveis de inteligência mais próximos aos níveis atuais?

A melhor opção dependeria da natureza do desafio para o qual estaríamos nos preparando. Se o desafio fosse a resolução de um problema para o qual o aprendizado pela experiência fosse crucial, então a duração cronológica do período de preparação poderia ser um fator determinante, uma vez que o tempo é necessário para que a experiência necessária se acumule. Com o que um desafio como esse se pareceria? Hipoteticamente, podemos usar como exemplo novas tecnologias armamentistas cuja criação, em algum momento no futuro, pudesse ser prevista e que, caso fossem usadas em alguma guerra subsequente, apresentassem uma chance de — digamos, uma em dez — gerar uma catástrofe existencial. Com um desafio dessa natureza, seria melhor para nós que o ritmo do desenvolvimento macroestrutural fosse mais lento, de maneira que nossa espécie tivesse mais tempo para se preparar antes da etapa crítica na qual essas novas tecnologias seriam inventadas. Esperaríamos que, durante o período de carência concedido pela desaceleração, nossa espécie pudesse ter aprendido como evitar guerras — que as relações internacionais pudessem ter chegado a um estágio que se assemelhasse às relações entre os países da União Europeia, os quais, tendo lutado ferozmente entre si durante séculos, hoje coexistem em paz e em relativa harmonia. Essa pacificação poderia ser lentamente construída através de diversos processos civilizatórios ou por meio de golpes subexistenciais, como em uma terapia de choque (por exemplo, pequenas conflagrações nucleares e os recuos e resoluções que estas poderiam gerar, culminando finalmente na criação das instituições globais necessárias para a abolição das guerras entre nações). Se esse tipo de aprendizado ou ajuste não pudesse ser acelerado por meio do aumento da inteligência, então o aprimoramento cognitivo não seria desejável e serviria apenas para queimar o pavio mais rapidamente.

No entanto, uma potencial explosão de inteligência poderia apresentar um desafio diferente. O problema do controle exige que tenhamos clarividência, capacidade de raciocínio e uma perspectiva teórica clara. Nesse caso, fica menos evidente de que maneira o crescimento de nossa experiência histórica poderia ajudar. Não é possível obter uma experiência direta de uma explosão de inteligência (não antes que seja tarde demais), e muitos elementos tornam o problema do controle algo singular, sem nenhum precedente histórico relevante. Por esses motivos, a quantidade de tempo que transcorrerá antes de uma explosão de inteligência pode, por si só, não ser relevante. Talvez importante seja, ao contrário, (a) a quantidade de progresso intelectual obtido em relação ao problema do controle até o momento da explosão de inteligência e (b) a quantidade de habilidades e inteligência disponíveis no momento da implementação das melhores soluções disponíveis (e improvisação de tudo o mais que esteja faltando).[498] Esse último fator deve, obviamente, responder positivamente a um aperfeiçoamento cognitivo. De que

maneira o aperfeiçoamento cognitivo influenciaria o fator (a) é uma questão, até certo ponto, mais sutil.

Suponha, conforme sugerido anteriormente, que o aperfeiçoamento cognitivo atuasse como um acelerador geral do desenvolvimento macroestrutural. Isso aceleraria a chegada de uma explosão de inteligência, reduzindo, então, a quantidade de tempo disponível para preparações e progresso na resolução do problema do controle. Normalmente, seria algo ruim. No entanto, se a aceleração do progresso intelectual fosse a única razão de haver menos tempo disponível para o progresso intelectual, então não seria necessária uma redução líquida na quantidade de progresso intelectual obtida até a ocorrência de uma explosão de inteligência.

Nesse ponto, o aperfeiçoamento cognitivo é aparentemente neutro em relação ao fator (a): a mesma quantidade de progresso intelectual obtida antes de uma explosão de inteligência sem o aceleramento — incluindo aqui o progresso em relação ao problema do controle — ainda poderia ser obtida, embora em um intervalo mais curto de tempo. Na realidade, entretanto, o aperfeiçoamento cognitivo pode acabar influindo positivamente em (a).

O aperfeiçoamento cognitivo poderia levar a um progresso maior em relação ao problema do controle até o momento da explosão de inteligência porque o progresso em relação ao problema do controle pode ser especialmente condicionado a valores extremos de desempenho intelectual — mais ainda do que aquele necessário para a criação da inteligência de máquina. O papel desempenhado pelo processo de tentativa e erro e pela acumulação de resultados experimentais parece bastante limitado em relação ao problema do controle, ao passo que o aprendizado experimental terá, provavelmente, um papel preponderante no desenvolvimento da inteligência artificial ou das emulações completas do cérebro. A proporção na qual o tempo é capaz de substituir a inteligência pode, desse modo, variar de acordo com a tarefa exercida, permitindo que o aperfeiçoamento cognitivo promova o progresso em relação ao problema do controle *em maior escala* do que promoveria o progresso em relação ao problema da criação de uma máquina inteligente.

Por outro lado, o aperfeiçoamento cognitivo promoveria especialmente o progresso em relação ao problema do controle porque a própria necessidade de tal progresso tem mais chance de ser apreciada por sociedades e indivíduos melhor capacitados cognitivamente. É necessário ter visão e raciocínio para entender por que o problema do controle é importante e fazer de sua solução uma prioridade.[499] Também é possível que seja necessária uma sagacidade incomum para encontrar maneiras promissoras de abordar um problema tão pouco familiar.

Com base nessas reflexões, podemos provisoriamente concluir que o aperfeiçoamento cognitivo é algo desejável, pelo menos na medida em que esteja concentrado nos riscos existenciais de uma explosão de inteligência. Linhas paralelas de raciocínio se aplicam a outros riscos existenciais ligados a desafios que demandam clarividência e raciocínio abstrato confiável (ao contrário, por exemplo, da adaptação incremental às mudanças experimentadas em um ambiente ou de um processo multigeracional de maturação cultural e criação de instituições).

Acoplamento de tecnologias

Suponha um cenário no qual se imagine que resolver o problema do controle para uma inteligência artificial seja algo muito difícil e que sua resolução no caso das emulações completas do cérebro se apresente como algo muito mais simples, sendo, portanto, preferível que a inteligência de máquina fosse alcançada por meio da emulação completa do cérebro. Mais adiante, discutiremos se as emulações completas do cérebro são ou não mais seguras do que a inteligência artificial. Mas, por enquanto, gostaríamos de assinalar que, embora aceitemos essa premissa, não significa que devemos promover tecnologias de emulação completa do cérebro. Uma razão, já discutida aqui, é que uma chegada tardia à superinteligência poderia nos dar mais tempo para aprimorar soluções ao problema do controle e para que um contexto mais favorável se desenvolvesse — e então, se estivéssemos confiantes de que a emulação completa do cérebro precederia a IA de qualquer forma, seria contraproducente continuar acelerando a chegada das emulações completas do cérebro.

Porém, embora fosse melhor que as emulações surgissem o quanto antes, *ainda assim* não deveríamos favorecer os avanços em direção à emulação completa do cérebro, pois é possível que esses avanços não resultem necessariamente em emulações completas do cérebro. Eles poderiam resultar, em vez disso, em uma inteligência artificial neuromórfica — uma forma de inteligência artificial que imita alguns aspectos da organização cortical, mas não replica fielmente as funcionalidades neurais com fidelidade suficiente para constituir uma emulação propriamente dita. Se — como há razões para acreditar — tal IA neuromórfica for pior do que outros tipos de IA que seriam criados por outros meios e se, ao promovermos a emulação completa do cérebro, fizéssemos com que a IA neuromórfica surgisse primeiro, então nossa busca por um desfecho supostamente *melhor* (emulação completa do cérebro) levaria ao *pior* desfecho (IA neuromórfica); por outro lado, se tivéssemos buscado o *segundo melhor* desfecho (IA sintética), seria possível, de fato, alcançar o *segundo melhor* desfecho (IA sintética).

Acabamos de descrever um exemplo (hipotético) do que poderíamos denominar "acoplamento de tecnologias".[500] Esse termo se refere a uma condição na qual duas tecnologias têm uma relação cronológica previsível, de modo que o desenvolvimento de uma dessas tecnologias tem uma forte tendência de levar ao desenvolvimento da outra, como um precursor necessário, uma aplicação óbvia e irresistível ou um passo subsequente. Acoplamentos de tecnologia devem ser levados em consideração quando usamos o princípio do desenvolvimento tecnológico diferencial: não é interessante acelerar o desenvolvimento de uma tecnologia desejável *Y* se a única maneira de obter *Y* for através do desenvolvimento de uma tecnologia precursora *X*, extremamente indesejável, ou se o desenvolvimento de *Y* produzisse, imediatamente, uma tecnologia relacionada extremamente indesejável *Z*. Antes de se casar, leve em consideração os seus futuros sogros.

No caso da emulação completa do cérebro, o grau de acoplamento de tecnologias é discutível. No capítulo 2, vimos que, embora a emulação completa do cérebro demande grandes avanços em diversas tecnologias que servem como base

para o seu desenvolvimento, não é necessário um avanço teórico significativo. Em particular, não é necessário compreender de que maneira a cognição humana funciona, mas apenas que sejamos capazes de construir modelos computacionais de pequenas partes do cérebro, como diferentes tipos de neurônios. Porém, à medida que estivéssemos desenvolvendo a habilidade de emular cérebros humanos, uma enormidade de dados neuroanatômicos seriam coletados e modelos funcionais de redes corticais seriam, com certeza, largamente aprimorados. Tais avanços, ao que tudo indica, parecem favoráveis ao surgimento de uma IA neuromórfica antes que fôssemos capazes de realizar efetivamente uma emulação completa do cérebro.[501] Historicamente, existem diversos exemplos de técnicas de IA inspiradas na neurociência e na biologia. (Por exemplo: o neurônio de McCulloch-Pitts, os percéptrons e outros tipos de neurônios e redes neurais artificiais inspirados em pesquisas neuroanatômicas; aprendizado por reforço inspirado na psicologia behaviorista; algoritmos genéticos inspirados na teoria da evolução; arquiteturas de subsunção e hierarquias perceptuais inspiradas nas teorias da ciência cognitiva a respeito do planejamento motor e da percepção sensorial; sistemas imunológicos artificiais inspirados nas teorias da imunologia; inteligência de enxame inspirada na ecologia das colônias de insetos e outros sistemas auto-organizáveis; e, em robótica, controle reativo e controle baseado em comportamento inspirados nos estudos de locomoção animal.) E talvez ainda mais significativo: existem muitas questões importantes relativas à IA que poderiam ser potencialmente respondidas com base em mais estudos sobre o cérebro. (Por exemplo: Como o cérebro armazena representações estruturadas na memória de trabalho e na memória de longo prazo? Como o problema da ligação é resolvido? Qual é o código neural? Como os conceitos são representados? Existe alguma unidade padrão de processamento cortical, como a coluna cortical, e, se esse for o caso, como ela é conectada internamente e de que maneira sua funcionalidade depende dessas conexões? Como essas colunas podem ser interligadas entre si e como elas são capazes de aprender?)

Em breve teremos mais a dizer sobre os perigos relativos da emulação completa do cérebro, da IA neuromórfica e da IA sintética, mas, por enquanto, podemos apontar um outro acoplamento relevante de tecnologias: aquele que envolve a emulação completa do cérebro e a IA. Ainda que um esforço em direção à emulação completa do cérebro realmente resulte em uma emulação completa do cérebro (e não em uma IA neuromórfica) e mesmo que o advento de emulações completas do cérebro pudesse ser conduzido de forma segura, um risco adicional ainda existiria: aquele associado a *uma segunda transição*, a qual seria uma transição da emulação completa do cérebro para a IA, que é a forma mais poderosa de inteligência de máquina.

Existem muitos outros tipos de acoplamento de tecnologias que poderiam ser considerados em uma análise mais aprofundada. Por exemplo, um esforço em direção à emulação completa do cérebro poderia desencadear avanços na neurociência como um todo.[502] Isso produziria diversos efeitos, como a aceleração de avanços em técnicas de detecção de mentiras, técnicas de manipulação neuropsicológicas, aperfeiçoamentos cognitivos e uma variedade de avanços na

medicina. Do mesmo modo, um esforço em direção ao aperfeiçoamento cognitivo poderia (dependendo do caminho específico que seria percorrido) repercutir em outras áreas, por exemplo, acelerando o desenvolvimento de métodos de seleção e engenharia genética, não apenas visando o aperfeiçoamento cognitivo, mas também modificações em outros aspectos.

Antecipando reações

É possível encontrar outra camada de complexidade estratégica se levarmos em conta que não existe um centro de controle mundial unificado, perfeitamente benevolente e racional, que simplesmente implemente aquilo que se descobriu ser a melhor opção. Qualquer abstração a respeito "do que deveria ser feito" deve assumir a forma de uma mensagem concreta, que deverá ser inserida na arena da realidade retórica e política. Ela será ignorada, mal compreendida, distorcida ou apropriada para fins conflitantes entre si; ela baterá e rebaterá como uma bola de pinball, causará uma série de ações e reações, levando a uma cascata de consequências cujo resultado não terá, necessariamente, qualquer relação com as intenções de quem as emitiu originalmente.

Um operador sofisticado poderia tentar antecipar esses tipos de efeito. Considere, por exemplo, o seguinte modelo de argumentação para o prosseguimento de uma pesquisa que objetiva desenvolver uma tecnologia perigosa *X*. (Um argumento que segue esse modelo pode ser encontrado na obra de Eric Drexler. No caso de Drexler, *X* = nanotecnologia molecular.)[503]

1. Os riscos de *X* são grandes.
2. Reduzir esses riscos demandará um período de preparação considerável.
3. Essa preparação considerável se iniciará apenas quando a perspectiva de *X* for levada a sério por amplos setores da sociedade.
4. Os amplos setores da sociedade somente levarão a perspectiva de *X* a sério quando passar a existir um esforço de pesquisa considerável para o desenvolvimento de *X*.
5. Quanto mais cedo um esforço de pesquisa considerável for iniciado, mais tempo levará para que *X* seja obtido (porque esse esforço se inicia a partir de um nível mais baixo de tecnologias preexistentes que possibilitam a obtenção de *X*).
6. Desse modo, quanto mais cedo um esforço considerável de pesquisa for iniciado, maior será o período em que as

preparações necessárias estarão ocorrendo, e maior será a redução dos riscos.

7. Sendo assim, um esforço considerável de pesquisa para o desenvolvimento de *X* deve ser iniciado imediatamente.

O que inicialmente parece uma razão para diminuir a velocidade ou parar completamente — os riscos grandes de *X* — acaba se tornando, com base nessa linha de raciocínio, uma razão para a conclusão oposta.

Segundo um tipo semelhante de argumento, deveríamos — de uma forma até insensível — ser gratos às catástrofes de pequena e média escala, já que elas nos alertam em relação às nossas vulnerabilidades e nos forçam a tomar precauções que reduzem a probabilidade de uma catástrofe existencial. A ideia é que catástrofes de pequena e média escala agem como uma inoculação, desafiando a civilização com uma ameaça à qual é relativamente possível sobreviver e estimulando uma resposta imunológica que prepararia o mundo para lidar com a versão existencial dessa ameaça.[504]

Esses argumentos de "force-os a reagir" advogam que devemos deixar algo ruim acontecer na esperança de que isso galvanize uma reação pública. Eles são mencionados aqui não com o intuito de endossá-los, mas apenas para introduzir a ideia do que denominaremos de "argumentos de antecipação". Tais argumentos sustentam que, ao tratar os outros como irracionais e manipular seus preconceitos e percepções errôneas, é possível obter uma resposta melhor do que aquela que teríamos se contássemos com a racionalidade de nossos interlocutores, apresentando nossos argumentos de forma honesta e direta.

Pode parecer impraticavelmente difícil utilizar o tipo de estratagema recomendado pelos argumentos de antecipação para alcançar objetivos globais de longo prazo. Como alguém poderia ser capaz de prever o destino final de uma mensagem após ela ter quicado e ricocheteado dentro da máquina de pinball que é a opinião pública? Fazê-lo poderia demandar uma previsão dos efeitos retóricos de uma miríade de constituintes, com variadas idiossincrasias e níveis flutuantes de influência por um longo período de tempo, durante o qual o sistema pode ser perturbado por eventos externos não antecipados enquanto a sua própria topologia passa por uma contínua reorganização endógena: certamente essa é uma tarefa impossível![505] No entanto, pode não ser necessário fazer uma previsão detalhada de toda a trajetória futura do sistema para identificar uma intervenção capaz de aumentar, razoavelmente, as chances de se obter um objetivo de longo prazo específico. Poderíamos, por exemplo, considerar detalhadamente apenas os efeitos que sejam relativamente de curto prazo e previsíveis, selecionando uma ação que pareça efetiva em relação a eles e modelando o comportamento do sistema para além desse horizonte de previsibilidade como um passeio aleatório.

Poderia existir, entretanto, um motivo moral para não enfatizar ou evitar a antecipação. Quando tentamos ser mais espertos do que os outros, entramos em uma disputa sem vencedores — ou em que há apenas perdedores, quando consideramos o tempo perdido e a energia que seria dissipada por essa prática,

além da probabilidade de que isso, em geral, tornaria difícil para qualquer um descobrir o que as outras pessoas realmente pensam e confiar em suas opiniões.[506] Uma utilização desenfreada de práticas de comunicação estratégica destruiria a sinceridade e deixaria a verdade desprotegida, lutando por conta própria em um mundo de raposas políticas traiçoeiras.

Caminhos e facilitadores

Devemos comemorar os avanços obtidos em hardware de computadores? E os avanços em direção à emulação completa do cérebro? Abordaremos essas duas questões a seguir.

Os efeitos dos avanços em hardware

Computadores mais rápidos facilitam a criação da inteligência de máquina. Um dos efeitos de acelerar os avanços em hardware seria, consequentemente, a antecipação do surgimento de máquinas inteligentes. Conforme discutido anteriormente, isso provavelmente seria algo ruim de acordo com a perspectiva impessoal, uma vez que reduziria o tempo disponível para a resolução do problema do controle e para que a humanidade atingisse um estágio civilizatório mais maduro. Mas essa, no entanto, não é uma questão resolvida. Uma vez que a superinteligência eliminaria muitos outros riscos existenciais, poderíamos preferir que seu desenvolvimento ocorresse mais cedo, caso os níveis de risco existencial se apresentassem elevados.[507]

A antecipação ou atraso do início de uma explosão de inteligência não é o único canal por meio do qual a taxa de avanço em hardware afetaria riscos existenciais. Outro canal está associado ao fato de que o hardware pode, em certa medida, ser um substituto para o software; desse modo, um hardware melhor poderia reduzir a quantidade mínima de habilidade necessária para programar uma IA embrionária. Computadores rápidos também podem encorajar a utilização de abordagens que dependam mais de técnicas de força bruta (como algoritmos genéticos e outros métodos de geração, avaliação e descarte) e menos de técnicas que demandam um entendimento mais profundo para serem utilizadas. Se as técnicas de força bruta resultarem em sistemas mais anárquicos e imprecisos, um cenário em que o problema do controle é mais difícil de ser resolvido do que no caso de sistemas mais precisamente elaborados e teoricamente controlados, essa seria uma outra maneira pela qual computadores rápidos poderiam aumentar os riscos existenciais.

Também devemos considerar que um avanço rápido em hardware aumenta a probabilidade de uma partida rápida. Quanto mais rápido for o avanço no estado da arte da indústria de semicondutores, menor será a quantidade de horas necessárias para que os programadores explorem as capacidades dos computadores, independentemente de seu nível de desempenho. Isso significa que é menos provável que uma explosão de inteligência seja iniciada nos níveis mais baixos de desempenho de hardware nos quais ela fosse viável. Desse modo, é *mais* provável que uma explosão de inteligência se inicie num momento em que o hardware tenha se desenvolvido significativamente além do nível mínimo no qual uma abordagem

bem-sucedida de programação pudesse se desenvolver. Existiria então um excedente de hardware quando finalmente ocorresse a partida. Conforme vimos no capítulo 4, o excedente de hardware é um dos principais fatores responsáveis pela redução da resistência durante uma partida. Assim, um progresso rápido em hardware tenderia a tornar mais rápida e explosiva a transição para a superinteligência.

Uma partida mais rápida por conta de um excedente de hardware pode afetar os riscos da transição de muitas maneiras. A mais óbvia delas é que uma partida mais rápida ofereceria menos oportunidades de resposta e execução de ajustes conforme ocorresse a transição, o que aumentaria os riscos. Uma outra consideração semelhante é que um excedente de hardware reduziria as chances de que uma IA embrionária, que estivesse se autoaperfeiçoando perigosamente, pudesse ser contida através da limitação de sua capacidade de adquirir hardware suficiente: quanto mais rápido for cada processador, menos processadores seriam necessários para que a IA fosse capaz de se converter mais rapidamente em uma superinteligência. Outro efeito do excedente de hardware é o nivelamento dos projetos grandes e pequenos, uma vez que haveria uma redução nas vantagens dos projetos maiores — a capacidade de custear computadores mais poderosos. Esse efeito também poderia contribuir para um aumento dos riscos existenciais caso projetos maiores sejam mais propensos a resolver o problema do controle e a buscar objetivos morais mais aceitáveis.[508]

Existem também algumas vantagens em uma partida mais rápida. Uma partida mais rápida aumentaria a probabilidade de formação de um *singleton*. Se o estabelecimento de um *singleton* for suficientemente importante para a resolução dos problemas de coordenação no pós-transição, poderia ser válido aceitar um risco maior durante a fase de explosão de inteligência para diminuir o risco de falhas de coordenação catastróficas posteriormente.

Os avanços na computação podem afetar o resultado de uma revolução de inteligência de máquina não apenas por desenrolarem um papel central na construção da inteligência de máquina, mas também por surtirem uma série de efeitos difusos na sociedade que ajudariam indiretamente a moldar as condições iniciais da explosão de inteligência. Atualmente, a internet, que demandou uma qualidade suficiente de hardware para permitir a produção de computadores pessoais em massa e a um baixo custo, influencia a atividade humana em diversas áreas, incluindo trabalhos em inteligência artificial e pesquisas a respeito do problema do controle. (Não fosse a internet, este livro poderia não ter sido escrito e você provavelmente não o teria encontrado.) Entretanto, o hardware já é bom o suficiente para um grande número de aplicações que poderiam facilitar a comunicação e deliberação humanas, e ainda não é muito claro se o ritmo do progresso nessas áreas depende fortemente da taxa de avanços em hardware.[509]

No geral, tudo indica que um progresso mais rápido em hardware e em computação é algo indesejável do ponto de vista de uma abordagem impessoal avaliativa. Essa conclusão provisória pode ser invalidada, por exemplo, caso as ameaças de outros riscos existenciais ou falhas de coordenação no pós-transição se mostrem extremamente altas. De qualquer maneira, parece pouco provável que

a velocidade dos avanços em hardware venha a fazer muita diferença. Sendo assim, nossos esforços em direção à melhoria das condições iniciais para uma explosão de inteligência deveriam estar concentrados em outros parâmetros.

Note que, ainda que não saibamos ao certo como influenciar algum parâmetro, pode ser útil determinar seu "sinal" (ou seja, determinar se um aumento ou diminuição nesse parâmetro seria desejável) como um passo anterior ao mapeamento do quadro estratégico geral. Poderíamos, futuramente, descobrir uma nova forma de influência que nos permitisse manipular esse parâmetro mais facilmente. Ou poderíamos descobrir que o sinal desse parâmetro possui correlação com o sinal de outro parâmetro que pudesse ser mais facilmente manipulado, e então nossa análise inicial poderia nos ajudar a decidir o que fazer com esse outro parâmetro.

As pesquisas em emulação completa do cérebro deveriam ser promovidas?

Quanto mais difícil se mostrar a resolução do problema do controle no caso de inteligências artificiais, mais tentador é promover o caminho da emulação completa do cérebro como uma alternativa menos arriscada. Existem, no entanto, diversas questões que devem ser analisadas antes que possamos chegar a um veredicto adequado a respeito do assunto.[510]

Primeiramente, existe o problema do acoplamento de tecnologias, discutido anteriormente. Demonstramos que um esforço para o desenvolvimento de uma emulação completa do cérebro poderia resultar, ao contrário, em uma IA neuromórfica, uma forma de inteligência de máquina que poderia ser particularmente insegura.

Mas, por uma questão de argumentação, partiremos do pressuposto de que uma emulação completa do cérebro (ECC) tenha sido desenvolvida. Ela seria mais segura do que uma IA? Essa é uma questão complicada. A ECC apresenta pelo menos três *supostas* vantagens: (i) suas características de desempenho seriam melhor compreendidas do que aquelas de uma IA; (ii) ela poderia herdar motivações humanas; e (iii) ela resultaria em uma partida lenta. Façamos um breve resumo de cada um desses pontos.

i

Parece plausível imaginar que seria mais fácil compreender as características de desempenho intelectual de uma emulação do que as de uma IA. Temos uma vasta experiência em relação aos pontos fortes e fracos da inteligência humana, mas nenhuma a respeito de uma inteligência artificial de nível humano. No entanto, entender o que um "retrato" de um intelecto humano digitalizado poderia ou não ser capaz de fazer não é o mesmo que compreender a maneira pela qual esse intelecto responderia às modificações que aumentariam seu desempenho. Por outro lado, um intelecto artificial poderia ser cuidadosamente desenvolvido para ser compreensível, tanto em relação às suas

predisposições estáticas quanto dinâmicas. Desse modo, ainda que uma emulação completa do cérebro possa ser mais previsível em relação ao seu desempenho intelectual do que uma IA genérica em estágio semelhante de desenvolvimento, não é óbvio se uma emulação completa do cérebro poderia ser dinamicamente mais previsível do que uma IA desenvolvida por programadores competentes e atentos às questões de segurança.

ii

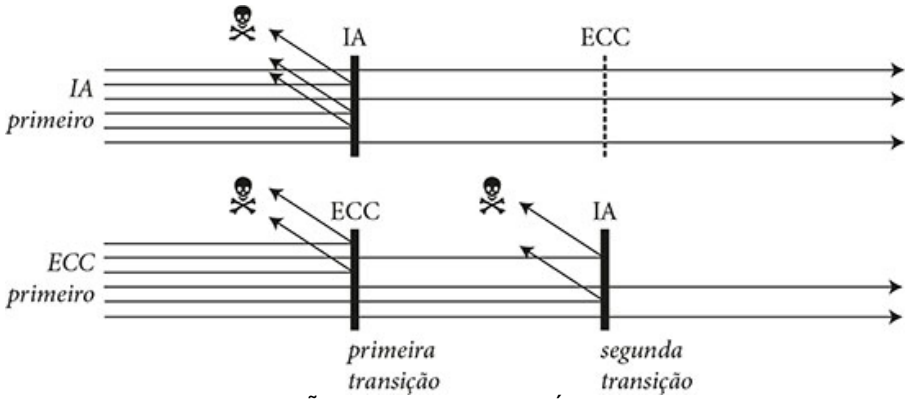
Por sua vez, a herança das motivações de seu modelo humano por parte da emulação é um fator que está longe de ser garantido. A captura de predisposições avaliativas humanas poderia demandar uma emulação de alta fidelidade. Ainda que as motivações de algum indivíduo *fossem* capturadas perfeitamente, não sabemos ao certo até que ponto isso seria seguro. Humanos podem ser pouco confiáveis, egoístas e cruéis. Mesmo que os modelos para as emulações fossem selecionados com base em suas virtudes excepcionais, pode ser muito difícil prever como alguém agiria quando transportado para circunstâncias radicalmente desconhecidas, com sua inteligência aumentada a níveis sobre-humanos e seduzido com uma oportunidade de dominar o mundo. É verdade que as emulações teriam, ao menos, uma chance maior de possuir motivações semelhantes às dos humanos (em oposição a valorizar apenas clipes de papel ou descobrir os dígitos de pi). Dependendo do ponto de vista de cada um em relação à natureza humana, isso poderia ou não ser confortante.[511]

iii

Não é tão evidente por que a emulação completa do cérebro deva resultar em uma partida mais lenta do que no caso da inteligência artificial. É possível esperar um menor efeito do excedente de hardware com a emulação completa do cérebro, pois ela seria menos computacionalmente eficiente do que as IAs. Igualmente, talvez um sistema de IA fosse capaz de absorver mais facilmente todo o poder computacional disponível e transformá-lo em um grande intelecto integrado, enquanto a emulação completa do cérebro prescindiria de uma superinteligência de qualidade e se sobressairia à humanidade apenas em velocidade de processamento e tamanho populacional. Se a emulação completa do cérebro de fato resultar em uma partida mais lenta, poderíamos ter alguns benefícios no sentido de aliviar o problema do controle. Uma partida mais lenta também aumentaria a probabilidade de ocorrência de um cenário multipolar. No entanto, ainda não temos certeza se um cenário multipolar seria algo desejável.

Existe uma outra complicação importante em relação à ideia de que seria mais seguro obter primeiro uma emulação completa do cérebro: o fato de que teríamos que lidar com uma *segunda transição*. Mesmo que a primeira forma de inteligência de máquina de nível humano seja baseada em uma emulação, ainda assim uma inteligência artificial poderia se desenvolver. Uma IA em sua forma madura tem vantagens relevantes em relação à ECC, o que faz da IA uma tecnologia muito mais poderosa.[512] Assim, uma IA madura tornaria a ECC obsoleta (a não ser para o propósito de preservar mentes humanas individuais), mas o inverso não é verdadeiro.

FIGURA 13



INTELIGÊNCIA ARTIFICIAL OU EMULAÇÃO COMPLETA DO CÉREBRO PRIMEIRO?. Em um cenário no qual a IA surgisse primeiro, existiria apenas uma transição que geraria um risco existencial. Em um cenário onde uma ECC surgisse primeiro, existiriam duas transições de risco, a primeira em relação à ECC e a segunda relacionada à IA. O total de risco existencial presente no cenário em que a ECC surge primeiro é a soma do risco dessas duas transições. No entanto, o risco de uma transição que envolva uma IA pode ser menor caso ela ocorra em um mundo onde as ECCs já tenham sido introduzidas satisfatoriamente.

Isso significa afirmar que, caso uma IA seja desenvolvida primeiro, teríamos, provavelmente, apenas uma única onda de explosão de inteligência. Por outro lado, se uma ECC for desenvolvida primeiro, provavelmente haveria duas ondas: primeiro o surgimento da ECC e, posteriormente, o surgimento da IA. O risco existencial total associado ao surgimento inicial de uma ECC equivale à *soma* do risco associado a uma primeira transição e do risco associado a uma segunda transição (caso sobrevivêssemos à primeira transição); ver figura 13.[513]

O quão mais segura seria uma transição para a IA em um mundo em que as ECCs estivessem presentes? Podemos considerar que essa transição para a IA seria menos explosiva caso ela ocorresse após o surgimento de algum tipo de inteligência de máquina. Emulações rodando a velocidades digitais e em números que poderiam exceder em muito a população humana biológica reduziriam o diferencial cognitivo, tornando mais fácil o controle da IA por parte das emulações. Essa consideração não é tão forte, uma vez que a lacuna entre uma IA e uma ECC poderia, ainda assim, ser considerável. Entretanto, se as simulações não forem apenas mais rápidas e mais numerosas, mas também, de algum modo, qualitativamente mais inteligentes que os humanos biológicos (ou, pelo menos, selecionadas no topo da distribuição de inteligência humana), então um cenário no

qual a ECC surgisse primeiro apresentaria vantagens similares àquelas de um cenário que envolvesse seres humanos aprimorados cognitivamente, conforme discutimos anteriormente.

Também devemos considerar que uma transição para a ECC aumentaria a vantagem do projeto líder. Consideremos um cenário onde o projeto líder estivesse seis meses à frente no desenvolvimento de tecnologias para a ECC em relação ao projeto na segunda colocação. Suponhamos que as primeiras emulações a serem criadas sejam cooperativas, focadas em segurança e pacientes. Se essas emulações rodassem em hardware rápido, elas poderiam passar éons subjetivos ponderando a respeito de como criar uma IA segura. Por exemplo, se elas rodassem a uma velocidade 100 mil vezes mais rápida que a do cérebro biológico e fossem capazes de trabalhar no problema do controle sem sofrer nenhum tipo de perturbação durante seis meses de tempo sideral, elas poderiam pensar a respeito do problema do controle ao longo de um período equivalente a cinquenta milênios antes que tivessem que competir com outras emulações. Caso houvesse hardware suficiente, elas seriam capazes de acelerar seu progresso criando uma miríade de cópias que trabalhariam de forma independente em subproblemas. Caso o projeto líder use essa vantagem de seis meses para formar um *singleton*, ele poderia dar à sua equipe de emulações responsável pela criação da IA uma quantidade ilimitada de tempo para trabalhar no problema do controle.[514]

No geral, tudo indica que o risco de uma transição para a IA seria menor caso uma ECC surgisse primeiro. No entanto, quando combinamos o risco residual da transição para a IA com o risco da transição anterior para a ECC, fica bem pouco claro de que maneira o risco existencial total em um cenário no qual uma ECC surgisse primeiro se compararia ao risco em um cenário onde uma IA surgisse primeiro. Um cenário no qual uma ECC surgisse primeiro poderia ser considerado mais atrativo apenas se formos muito pessimistas em relação à habilidade da humanidade biológica em lidar com uma transição para IA — levando em consideração que a natureza humana ou a civilização poderiam ter se aprimorado quando for chegada a hora de enfrentar esse desafio.

A fim de concluir se deveríamos ou não promover as tecnologias para a emulação completa do cérebro, é necessário que mais algumas questões importantes sejam colocadas na balança. A mais importante delas é o acoplamento de tecnologias, mencionado anteriormente: um esforço em direção à ECC poderia produzir, em vez de emulações completas do cérebro, uma IA neuromórfica. Esse é um dos motivos para não incentivarmos esforços em direção à ECC.[515] Existem, sem dúvida, *alguns* projetos sintéticos de IA que seriam menos seguros do que *alguns* projetos neuromórficos. Em média, entretanto, os projetos neuromórficos parecem menos seguros. Uma das bases para essa afirmação é o fato de que a imitação pode funcionar como um substituto para a compreensão. Para construir algo do zero, geralmente é necessária uma compreensão razoavelmente boa a respeito do possível funcionamento do sistema. Tal compreensão pode não ser necessária se o intuito for meramente copiar aspectos de um sistema preexistente. A emulação completa do cérebro se baseia na cópia em larga escala da biologia, que pode não demandar uma compreensão computacional detalhada da cognição do sistema

como um todo (no entanto, um grande entendimento em relação aos componentes desse sistema seria, sem dúvida, necessário.) A IA neuromórfica tem esse aspecto em comum com a emulação completa do cérebro: ela seria obtida através da junção de peças plagiadas da biologia, dispensando que seus criadores possuam um entendimento matemático profundo a respeito do funcionamento do sistema. No entanto, em outro aspecto, uma IA neuromórfica seria *diferente* de uma emulação completa do cérebro: ela não teria motivações humanas de antemão.[516] Essa consideração argumenta contra esforços em direção à emulação completa do cérebro, na medida em que eles provavelmente produziriam uma IA neuromórfica.

Também devemos colocar na balança o fato de que a ECC estaria mais propensa a nos avisar previamente a respeito de seu surgimento. No caso de uma IA, é sempre possível que alguém faça uma descoberta conceitual inesperada. A ECC, ao contrário, demanda uma série de etapas laboriosas — escâneres de alta resolução, softwares de processamento de imagens e modelagem neural detalhada. Desse modo, podemos ter bastante certeza de que a ECC não é uma tecnologia iminente (não surgirá, digamos, nos próximos quinze ou vinte anos). Isso significa dizer que esforços para acelerar a ECC fariam diferença apenas em cenários nos quais a inteligência de máquina seja desenvolvida comparativamente tarde. Isso pode tornar atrativos os investimentos em ECC para alguém cujo objetivo seja impedir outros riscos existenciais por meio da explosão de inteligência, mas que tenha receio de apoiar pesquisas em IA por conta da possibilidade de desencadear uma explosão de inteligência prematuramente, antes que o problema do controle tenha sido solucionado. No entanto, atualmente, a incerteza em relação às escalas de tempo relevantes é provavelmente muito grande para permitir que essa consideração tenha muito peso.[517]

Desse modo, uma estratégia de promoção da ECC seria mais atrativa se (a) formos muito pessimistas em relação à capacidade humana de resolver o problema do controle no caso da IA, (b) não tivermos muita preocupação em relação à IA neuromórfica, aos resultados multipolares ou aos riscos associados a uma segunda transição, (c) acharmos que a diferença de tempo entre o surgimento de uma ECC e de uma IA será pequena e (d) preferirmos que uma superinteligência não se desenvolva nem muito tarde, nem muito cedo.

A perspectiva do indivíduo-afetado favorece uma maior velocidade

Temo que o comentário feito por “washbash” em um blog reflita a opinião de muitas pessoas:

> Instintivamente, eu penso: vá mais rápido. Não porque eu creia que isso seja melhor para o mundo. Por que eu deveria me importar com o planeta quando eu já estiver morto e enterrado? Que se dane, eu quero que tudo ande rápido! Isso aumenta as minhas chances de vivenciar um futuro mais avançado tecnologicamente.[518]

Do ponto de vista do indivíduo-afetado, temos mais razões para apressar de todas as maneiras possíveis o surgimento de tecnologias radicais que poderiam representar riscos existenciais. Isso se dá porque o resultado padrão é que quase todas as pessoas existentes hoje em dia já estarão mortas dentro de um século.

A defesa da aceleração é especialmente forte em relação a tecnologias que seriam capazes de aumentar nossa expectativa de vida e, consequentemente, incrementar a proporção da população que ainda estaria viva quando ocorresse uma explosão de inteligência. Se a revolução de inteligência de máquina ocorrer de forma satisfatória, a superinteligência dela resultante poderia quase certamente criar meios de prolongar indefinidamente a vida dos humanos existentes, não apenas mantendo-os vivos, mas restabelecendo saúde e vigor de suas juventudes e aumentando suas capacidades para muito além do que hoje consideramos como limites humanos; ou poderia ajudá-los a eliminar completamente sua mortalidade, transferindo suas mentes para um substrato digital, onde seus espíritos livres teriam uma enorme sensação de bem-estar em seus novos corpos digitais. Em relação às tecnologias que não prometem salvar vidas, a defesa de uma aceleração dos avanços tecnológicos é mais fraca, mas talvez ainda suficientemente ancorada na esperança de que essas tecnologias seriam capazes de elevar os padrões de vida.[519]

Seguindo a mesma linha de raciocínio, a perspectiva do indivíduo-afetado favorece muitas outras inovações tecnológicas arriscadas que prometem acelerar o advento de uma explosão de inteligência, mesmo nos casos em que essas inovações se mostram desfavoráveis a partir da perspectiva impessoal. Tais inovações poderiam diminuir o período de escuridão pelo qual teríamos que passar se quiséssemos ver o nascer do dia em uma era pós-humana. Do ponto de vista do indivíduo-afetado, um progresso mais rápido em hardware parece algo desejável, bem como um progresso mais veloz em direção à ECC. Qualquer efeito adverso relacionado ao risco existencial seria provavelmente compensado pelo benefício pessoal relacionado à maior probabilidade de que uma explosão de inteligência aconteça durante o período de vida dos indivíduos atualmente existentes.[520]

Colaboração

Um parâmetro importante é o grau de coordenação e colaboração que o mundo será capaz de atribuir ao desenvolvimento da inteligência de máquina. A colaboração poderia trazer muitos benefícios. Analisaremos agora como esse parâmetro seria capaz de afetar o resultado e quais mecanismos poderiam ser utilizados para aumentar a extensão e a intensidade da colaboração.

A dinâmica competitiva e seus perigos

Uma dinâmica competitiva existe quando um projeto teme ser dominado por outro. Isso não requer, necessariamente, a existência de múltiplos projetos. Uma situação na qual um único projeto poderia exibir uma dinâmica competitiva ocorre quando não há uma certeza em relação à inexistência de competidores. Os Aliados provavelmente jamais teriam desenvolvido a bomba atômica na velocidade em que

o fizeram se eles não tivessem acreditado (erroneamente) que os alemães poderiam estar próximos de alcançar esse mesmo objetivo.

A intensidade de uma dinâmica competitiva (ou seja, o quão priorizada é, por parte dos competidores, a velocidade em detrimento da segurança) depende de diversos fatores, como o quão próximos os competidores estão uns dos outros, a importância relativa entre capacidade e sorte, o número de competidores, se os projetos que competem entre si se utilizam de abordagens diferentes ou não e o grau de compartilhamento dos mesmos objetivos entre os projetos. As crenças dos competidores em relação a esses fatores também são relevantes. (Ver quadro 13.)

Durante o desenvolvimento de uma superinteligência de máquina, tudo indica que haverá pelo menos uma dinâmica competitiva leve, e é possível que exista uma dinâmica competitiva severa. A dinâmica competitiva tem consequências importantes em relação à maneira como deveria ser pensado o desafio estratégico lançado pela possibilidade de uma explosão de inteligência.

A dinâmica competitiva poderia estimular os projetos a se moverem mais rapidamente em direção a uma superinteligência e, consequentemente, eles reduziriam investimentos na resolução do problema do controle. Outros efeitos prejudiciais da dinâmica competitiva também são possíveis, tais como hostilidades diretas entre os competidores. Suponha que duas nações estejam competindo para desenvolver a primeira superinteligência e que uma delas esteja, aparentemente, na liderança. Em um cenário no qual o vencedor ficará com todos os benefícios, o projeto que está em desvantagem pode ficar tentado a lançar um ataque desesperado contra seu rival, em vez de esperar passivamente pela derrota. O projeto líder poderia antecipar essa possibilidade e ser tentado a atacar preventivamente. Se os antagonistas forem nações poderosas, a luta poderia ser sangrenta.[521] (Um "ataque cirúrgico" contra o projeto de IA rival desencadearia um confronto maior e poderia, de qualquer maneira, não ser algo factível caso o país do projeto rival tenha tomado as devidas medidas de precaução.)[522]

QUADRO 13

Um nivelamento por baixo arriscado[523]

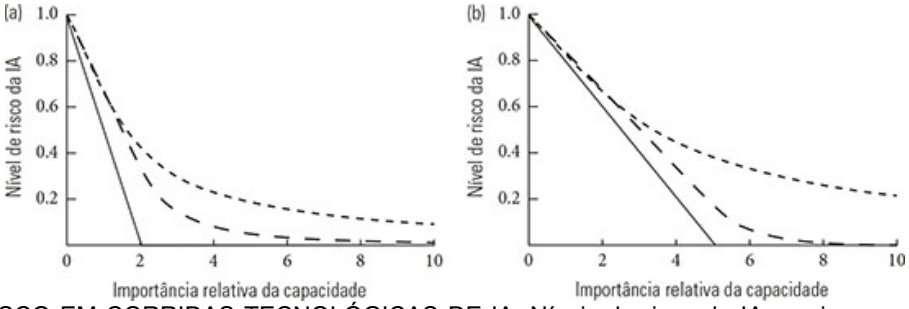
Considere uma hipotética corrida armamentista de IA na qual vários grupos rivais competem para desenvolver uma IA superinteligente.[524] Cada grupo decide o quanto será investido em segurança — sabendo que os recursos gastos no desenvolvimento de medidas de segurança são recursos não utilizados no desenvolvimento da IA. Na ausência de um acordo entre todos os competidores (algo que poderia ser difícil alcançar em virtude das dificuldades de negociação e verificação), poderia ocorrer um nivelamento por baixo arriscado, levando cada grupo a tomar apenas precauções mínimas de segurança.

Podemos modelar o desempenho de cada grupo como uma função de suas capacidades (medindo suas habilidades brutas e sua sorte) e impor uma penalidade que correspondesse ao custo de suas precauções de segurança. O grupo com o maior desempenho construirá a primeira IA. O grau de risco dessa IA é determinado pela quantidade de investimentos em segurança realizados por seus criadores. No pior cenário possível, todos os grupos apresentam níveis iguais de capacidade. Nesse caso, o vencedor será determinado exclusivamente pelo investimento em segurança: o grupo que tomar as menores precauções de segurança será o vencedor. O equilíbrio de Nash para esse jogo é que nenhum dos grupos gere qualquer gasto em segurança. No mundo real, tal situação poderia se dar por meio de um efeito que chamamos de *catraca de riscos*: um grupo, receoso de ficar para trás, passa a assumir mais risco para se equiparar a seus competidores — que respondem do mesmo modo, até que o nível máximo de risco seja alcançado.

Capacidade versus risco

A situação muda quando há divergências de capacidade entre os competidores. Conforme as diferenças de capacidade se tornem mais importantes do que os custos com precauções de segurança, o efeito da catraca de risco se enfraquece: temos, então, um incentivo menor para que um risco extra seja assumido caso essa decisão apresente uma probabilidade baixa de alterar a ordem da disputa. Essa questão está ilustrada para diversos cenários na figura 14, que mostra como o nível de risco de uma IA depende da importância da capacidade. Os investimentos em segurança variam de 1 (resultando em uma IA perfeitamente segura) 0 (uma IA completamente insegura). O eixo *x* representa a importância relativa da capacidade versus o investimento em segurança para determinar a velocidade do progresso de um grupo em direção à obtenção da IA. (Em 0,5, o nível de investimento em segurança é duas vezes mais importante do que a capacidade; em 1, os dois são iguais; em 2, o nível de capacidade é duas vezes mais importante que o nível de segurança e assim por diante.) O eixo *y* representa o nível de risco da IA (a fração esperada da utilidade máxima que será obtida pelo vencedor da disputa).

É possível notar que, em qualquer um dos cenários, a periculosidade da IA resultante é máxima quando a capacidade não possui um papel central, decrescendo gradualmente conforme a capacidade cresce em importância.



NÍVEIS DE RISCO EM CORRIDAS TECNOLÓGICAS DE IA. Níveis de risco de IAs perigosas em um modelo simples de uma corrida tecnológica envolvendo tanto (a) dois grupos ou (b) cinco grupos, representados em relação à importância relativa da capacidade (em vez do investimento em segurança) para determinar que projeto vencerá a disputa. Os gráficos mostram três cenários com níveis distintos de informação: sem informação a respeito da capacidade (linha reta), informação privada a respeito da capacidade (linha tracejada) e informação plena a respeito da capacidade (linha pontuada).

Objetivos compatíveis

Outra maneira de reduzir os riscos seria conceder aos grupos uma participação no sucesso dos outros grupos, independentemente de quem for o vencedor. Se os competidores estiverem convencidos de que chegar em segundo lugar significa a perda total de tudo aquilo que lhes é importante, eles assumirão qualquer tipo de risco necessário para ultrapassar seus rivais. Por outro lado, os grupos investirão mais em segurança se não for imprescindível vencer a disputa. Isso sugere que devemos incentivar diversas formas de investimento cruzado.

O número de competidores

Quanto maior for o número de grupos competidores, mais perigosa se torna a disputa: como cada grupo tem uma chance menor de ser o vencedor, maior será a chance de que esses grupos deixem a cautela de lado. Isso pode ser observado quando comparamos a figura 14a (dois grupos) com a figura 14b (cinco grupos). Em todos os cenários, o risco seria reduzido se os grupos se unissem em um número menor de coalizões competindo entre si.

A maldição do excesso de informação

É vantajoso que os grupos conheçam suas posições na disputa (reconhecendo quais seriam seus níveis de capacidade, por exemplo)? Aqui, fatores opostos entram em cena. É desejável que um líder saiba que está na liderança (de modo que ele esteja ciente de que possui alguma margem para tomar medidas de segurança adicionais). No entanto, não é desejável que um retardatário saiba que ele está em desvantagem (uma vez que isso confirmaria que as medidas de segurança precisam ser deixadas de lado para que possa haver alguma esperança de recuperação). Intuitivamente, é possível ter a impressão de que essa escolha poderia ser feita de ambas as formas, mas os modelos são inequívocos: informação (em média) é algo ruim.[2] As figuras 14a e 14b apresentam, cada uma, três cenários: as linhas retas correspondem a situações nas quais nenhum grupo possui informações a respeito dos níveis de capacidade, incluindo o seu próprio nível. As linhas tracejadas mostram situações onde cada grupo conhece apenas seu próprio nível de capacidade. (Nesses casos, um grupo assume riscos adicionais apenas caso seu nível de capacidade se apresente baixo.) E as linhas pontilhadas mostram o que acontece quando todos os grupos têm conhecimento a respeito do nível de capacidade de cada um. (Maiores riscos são assumidos se os níveis de capacidade forem semelhantes entre si.) A dinâmica competitiva se torna pior a cada aumento no nível de informação.

Cenários nos quais os desenvolvedores rivais não sejam nações, mas entidades menores, como laboratórios industriais ou equipes acadêmicas, provavelmente gerariam menos destruição direta resultante de conflitos. No entanto, as consequências de uma dinâmica competitiva seriam, de modo geral, quase tão graves. Isso se dá porque a parcela principal do dano esperado de uma dinâmica competitiva não advém da destruição gerada por eventuais batalhas e combates, mas da diminuição nos níveis de precaução. Conforme já discutimos, uma dinâmica competitiva causaria uma redução nos investimentos em segurança, e os conflitos, mesmo que não violentos, tenderiam a limitar as oportunidades de colaboração, uma vez que os projetos provavelmente teriam menos propensão a compartilhar ideias para resolver o problema do controle em meio a um clima de desconfiança e hostilidade.[525]

Sobre os benefícios da colaboração

A colaboração, desse modo, oferece muitos benefícios. Essa estratégia reduz a pressa no desenvolvimento de uma inteligência de máquina. Permite um maior investimento em segurança. Evita conflitos violentos e facilita o compartilhamento de ideias a respeito da resolução do problema do controle. Podemos, ainda, adicionar mais um benefício: a colaboração teria uma tendência de produzir cenários nos quais os frutos de uma explosão de inteligência bem controlada seriam distribuídos de forma mais igualitária.

O fato de que colaborações mais amplas resultariam em uma divisão mais ampla dos ganhos não é algo axiomático. Em princípio, um projeto pequeno, coordenado por um altruísta, poderia conduzir a um cenário no qual os benefícios seriam divididos igualmente ou de forma equilibrada entre todos os seres que fossem

moralmente consideráveis. Porém, existem diversas razões para supor que colaborações mais amplas, envolvendo um grande número de patrocinadores, seriam (em média) superiores em termos de distribuição dos benefícios. Uma razão para tanto é que os patrocinadores teriam, presumivelmente, preferência por um cenário no qual eles mesmos recebessem (pelo menos) uma parcela justa dos ganhos. Desse modo, uma colaboração ampla permitiria que relativamente muitos indivíduos recebessem, ao menos, uma parcela justa dos ganhos, considerando aqui que o projeto fosse bem-sucedido. Outra razão é que uma colaboração ampla, ao que tudo indica, teria uma maior probabilidade de beneficiar pessoas que não estivessem incluídas nessa colaboração. Uma colaboração mais ampla teria mais membros e, desse modo, um número maior de pessoas não envolvidas diretamente também receberia benefícios, pois eles teriam mais chance de possuir ligações pessoais com alguém que estivesse envolvido e estivesse preocupado com seus interesses. Uma colaboração mais ampla também tem maior possibilidade de incluir pelo menos algum altruísta que deseje o benefício de todos. Além disso, uma colaboração mais ampla teria mais chances de operar sob supervisão pública, o que poderia reduzir o risco de que todos os benefícios fossem capturados por um pequeno grupo de programadores ou de investidores privados.[526] Note também que quanto maior for a colaboração que tenha obtido sucesso, menor são os custos de expansão dos seus benefícios para todos aqueles que não estejam envolvidos no processo. (Por exemplo, se 90% de todas as pessoas já fizessem parte da colaboração, o custo para fazer com que todo o restante da população alcançasse o mesmo nível dos membros envolvidos não seria maior do que 10% do patrimônio de cada membro.)

Sendo assim, é plausível pensar que colaborações mais amplas tenderiam a gerar uma distribuição igualmente mais ampla dos ganhos (embora *alguns* projetos com poucos patrocinadores também possam ter objetivos excelentes em termos de distribuição dos ganhos). Mas por que uma distribuição ampla dos ganhos é algo desejável?

Existem razões tanto morais quanto prudenciais para se favorecer cenários nos quais todos recebam uma parte dos ganhos. Não entraremos nas questões de cunho moral; notem apenas que elas não precisam estar embasadas em nenhum princípio igualitário. Poderíamos defendê-las, por exemplo, com base em princípios de justiça. Um projeto responsável pela criação de uma superinteligência de máquina impõe uma externalidade global. Todos os habitantes do planeta seriam colocados em risco, incluindo aqueles que não concordassem em ter suas vidas ou as de seus entes queridos arriscadas dessa maneira. Como todos compartilhariam dos mesmos riscos, parece ser um requerimento mínimo de justiça que todos também compartilhem de qualquer ganho eventual.

O fato de que o total (esperado) de benefícios parece maior em cenários colaborativos é outra razão importante que torna tais cenários moralmente preferíveis.

No caso da prudência, há dois motivos principais que favoreceriam uma distribuição ampla dos ganhos. O primeiro é que uma distribuição ampla deve promover a colaboração, diminuindo as consequências negativas de uma dinâmica

competitiva. Haveria menos incentivo para uma disputa de qual seria o primeiro projeto a desenvolver uma superinteligência se todos, por fim, se beneficiassem igualmente com o sucesso de qualquer projeto. Os patrocinadores de um projeto em particular também poderiam se beneficiar caso sinalizassem, efetivamente, seu compromisso em distribuir de forma universal os benefícios obtidos por seu projeto, pois um projeto garantidamente altruísta atrairia mais apoiadores e menos inimigos.[527]

O outro motivo baseado na prudência por trás da defesa de uma ampla distribuição de ganhos está ligado ao fato de que os agentes são avessos a riscos ou possuem funções de utilidade sublineares em relação aos recursos. O fator central aqui é a enormidade de recursos em potencial. Considerando que o universo observável esteja tão inabitado quanto parece, ele possui mais do que uma galáxia disponível para cada ser humano vivo. A maioria das pessoas preferiria a garantia de acesso aos recursos existentes em uma galáxia do que um bilhete de loteria oferecendo a chance de um em 1 bilhão de ganhar 1 bilhão de galáxias.[528] Em virtude do tamanho astronômico da herança cósmica da humanidade, tudo indica que o interesse próprio deveria favorecer, geralmente, acordos que garantissem a cada pessoa uma parte dos recursos, ainda que essa parcela corresponda a uma pequena fração dos recursos totais. O mais importante é não ser deixado de fora quando tal prosperidade exorbitante estiver iminente.

Esse argumento a respeito da enormidade dos recursos pressupõe que essas preferências poderiam se satisfazer com uma determinada quantidade de recursos.[529] Mas essa não é uma suposição que, necessariamente, se sustenta. Por exemplo, diversas teorias éticas proeminentes — incluindo especialmente as teorias consequencialistas agregativas — correspondem a funções de utilidade que são neutras em relação aos riscos e lineares em relação aos recursos. 1 bilhão de galáxias poderiam ser usadas para gerar 1 bilhão de vezes mais vidas felizes do que uma única galáxia. Desse modo, para um utilitarista, elas valeriam 1 bilhão de vezes mais.[530] Funções de preferência humanas egoístas comuns, no entanto, parecem se satisfazer relativamente com recursos.

Esse último argumento deve ser acompanhado por duas importantes considerações. A primeira está ligada ao fato de que muitas pessoas se importam demasiadamente com a sua posição social. Se diversos agentes desejam figurar no topo da lista da *Forbes*, não haverá uma quantidade suficiente de recursos para que todos estejam plenamente satisfeitos.

A segunda consideração diz respeito ao fato que de que as bases tecnológicas do pós-transição permitiriam que os recursos materiais fossem convertidos em uma gama sem precedentes de produtos, incluindo alguns bens que atualmente não estão disponíveis, independentemente de seu preço, embora sejam muito valorizados pelos seres humanos. Um bilionário não vive mil vezes mais do que um milionário. Todavia, na era das mentes digitais, um bilionário poderia arcar com os custos de acesso a um poder computacional mil vezes maior e, dessa maneira, aproveitaria uma vida subjetiva mil vezes mais longa. Capacidade mental também é algo que poderia estar à venda. Num cenário em que o capital econômico pudesse ser convertido em bens vitais a uma taxa constante, mesmo no caso de altos níveis

de riqueza, a existência de uma ganância sem limites poderia fazer mais sentido do que na atualidade, quando os mais abastados (ao menos aqueles que não possuem nenhuma vocação filantrópica) estão limitados a gastar suas riquezas em aviões, barcos, coleções de arte ou na aquisição de uma quarta ou quinta residência.

Isso significa que um egoísta deveria ser neutro a riscos no que diz respeito à sua herança de recursos no pós-transição? Não necessariamente. Recursos físicos podem não ser conversíveis em mais tempo de vida ou em desempenho mental a níveis arbitrários. Se uma vida tiver que ser vivida sequencialmente, de modo que os momentos de observação possam lembrar eventos anteriores e sofrerem influências de escolhas passadas, então a vida de uma mente digital não poderia ser estendida arbitrariamente sem a utilização de um número cada vez maior de operações computacionais *sequenciais*. Mas a física limita a extensão na qual os recursos podem ser transformados em operações computacionais sequenciais.[531] Os limites da computação sequencial também podem restringir alguns aspectos do desempenho cognitivo, levando-o a aumentar de forma radicalmente sublinear a partir de uma quantidade relativamente modesta de recursos. Além disso, não está claro se um egoísta poderia ou não ser neutro em relação aos riscos, mesmo no que diz respeito a métricas de resultado bastante relevantes normativamente, tais como o número de anos subjetivos de vida ajustados de acordo com a qualidade de vida. Entre a possibilidade de garantia de 2 mil anos extras de vida e uma chance em dez de 30 mil anos extras de vida, acredito que a maioria das pessoas selecionaria a primeira opção (ainda que estivesse estipulado que cada ano de vida possuiria a mesma qualidade).[532]

Na realidade, a afirmação de que uma distribuição ampla dos ganhos seria algo prudente depende, presumivelmente, do sujeito e da situação. Contudo, de uma maneira geral, as pessoas teriam mais chances de obter (quase tudo) o que elas desejassem se encontrássemos uma maneira de alcançar uma ampla distribuição dos ganhos — e isso vale mesmo antes de levarmos em consideração que um compromisso em relação a uma distribuição mais ampla tenderia a promover a colaboração, aumentando, desse modo, as chances de se evitar uma catástrofe existencial. Sendo assim, favorecer uma distribuição ampla parece não ser apenas uma obrigação moral, mas também algo sensatamente aconselhável.

Há ainda outro conjunto de consequências ligadas à colaboração ao qual deveríamos dar pelo menos alguma atenção: a possibilidade de que uma colaboração no pré-transição influencie o nível de colaboração no pós-transição. Suponha que a humanidade tenha resolvido o problema do controle. (Caso o problema do controle ainda não tivesse sido resolvido, a quantidade de colaboração existente no pós-transição poderia importar muito pouco.) Há dois casos que devemos considerar. O primeiro é o cenário em que uma explosão de inteligência *não* cria uma dinâmica na qual o vencedor levará todos os ganhos (presumivelmente porque a partida seria relativamente lenta). Nesse caso, se a colaboração na pré-transição tiver qualquer efeito sistemático na colaboração no pós-transição, é possível que o efeito seja positivo, tendendo a promover uma colaboração subsequente. As relações colaborativas originais poderiam perdurar para além da transição; além disso, a colaboração no pré-transição poderia oferecer

maiores oportunidades para que as pessoas guiassem os desenvolvimentos em direções mais desejáveis (e presumivelmente mais colaborativas) no pós-transição.

No segundo caso, a natureza da explosão de inteligência encorajaria uma dinâmica na qual o vencedor poderia levar todos os ganhos (presumivelmente porque a partida seria rápida). Então, caso não haja uma colaboração extensiva antes da partida, é bem provável que surja um *singleton* — um único projeto passaria sozinho pela transição e em algum ponto obteria uma vantagem estratégica decisiva combinada com a superinteligência. Um *singleton* é, por definição, uma ordem social altamente colaborativa.[533] A ausência de colaboração extensiva no pré-transição conduziria, desse modo, a um nível extremo de colaboração no pós-transição. Em contrapartida, um nível um pouco mais alto de colaboração nos estágios imediatamente anteriores a uma explosão de inteligência permitiria possibilidades mais variadas de resultados. Projetos que colaborassem entre si poderiam sincronizar seu crescimento de modo a assegurar que a transição ocorresse ao mesmo tempo, sem que nenhum dos projetos viesse a obter uma vantagem estratégica decisiva. Ou diferentes grupos de patrocinadores poderiam juntar seus esforços em um único projeto ao mesmo tempo, recusando-se a permitir que esse projeto formasse um *singleton*. Por exemplo, podemos imaginar um consórcio de nações criando um projeto científico coletivo para desenvolver uma superinteligência de máquina, mas proibindo que esse projeto evoluísse a ponto de se tornar uma espécie de ONU reforçada, preferindo, em vez disso, manter a ordem mundial faccionada como era antes.

Desse modo, especialmente no caso de uma partida rápida, existe a possibilidade de que uma grande colaboração no pré-transição resultasse em uma menor colaboração no pós-transição. No entanto, na medida em que entidades colaborativas fossem capazes de moldar o resultado, seria possível permitir o aparecimento ou a continuidade de um cenário sem colaboração apenas se essas entidades pudessem prever que nenhuma consequência catastrófica estaria associada a um cenário do tipo no pós-transição. Portanto, cenários nos quais a colaboração no pré-transição conduzisse a uma redução da colaboração no pós-transição poderiam ser, principalmente, aqueles em que uma colaboração reduzida no pós-transição viesse a ser inócua.

Em geral, um grau mais acentuado de colaboração no pós-transição parece ser algo desejável. Isso reduziria o risco de uma dinâmica distópica na qual a competição econômica e uma rápida expansão da população levaria a uma condição malthusiana, ou a uma dinâmica na qual a seleção evolutiva erodisse os valores humanos e selecionasse formas não eudemônicas, ou em que potências rivais sofressem outros tipos de falhas de coordenação, como guerras e corridas tecnológicas. Esse último aspecto — a perspectiva de corridas tecnológicas — pode ser particularmente problemático caso ocorra uma transição para uma forma intermediária de inteligência de máquina (emulação completa do cérebro), o que criaria uma nova dinâmica competitiva que poderia prejudicar as chances de solução do problema do controle antes da segunda transição subsequente para uma forma mais avançada de inteligência de máquina (inteligência artificial).

Descrevemos anteriormente de que maneira a colaboração poderia reduzir conflitos nas fases imediatamente anteriores a uma explosão de inteligência, aumentando as chances de solução do problema do controle e aprimorando tanto a legitimidade moral quanto a conveniência em termos de prudência da alocação dos recursos resultantes. A esses benefícios associados à colaboração, é possível adicionar mais um: uma colaboração mais ampla no pré-transição poderia ajudar a resolver importantes problemas de coordenação na era do pós-transição.

Trabalhando em conjunto

Dependendo da dimensão das entidades colaborativas em questão, a colaboração pode assumir diversas formas. Em escalas menores, equipes individuas de IA que acreditassem estar em competição umas com as outras poderiam optar por uma junção de esforços.[534] Corporações poderiam se fundir ou fazer investimentos cruzados. Em uma escala maior, nações poderiam se unir em torno de um grande projeto internacional. Existem precedentes de colaborações internacionais de larga escala em ciência e tecnologia (como o a Organização Europeia para a Pesquisa Nuclear [CERN], o Projeto Genoma Humano e a Estação Espacial Internacional), porém um projeto internacional voltado ao desenvolvimento de uma superinteligência segura enfrentaria um desafio de ordem diferente em virtude das implicações de segurança envolvidas. Esse projeto não poderia ser constituído como uma colaboração acadêmica aberta, mas como um empreendimento fechado e altamente controlado. Provavelmente os cientistas que dele participassem precisariam se manter fisicamente isolados e impedidos de se comunicar com o restante do mundo durante a duração do projeto, a não ser por um único canal de comunicação que seria cuidadosamente monitorado. O nível de segurança necessário para tanto talvez fosse algo praticamente inalcançável nos dias atuais, mas avanços em detecção de mentiras e tecnologia de vigilância poderiam tornar essa empresa factível no fim deste século. Também vale a pena ter em mente que uma colaboração ampla não significa necessariamente que muitos pesquisadores estariam envolvidos no projeto, mas simplesmente que muitas pessoas teriam direito a voz em relação aos objetivos propostos. Em princípio, um projeto poderia envolver uma colaboração extrema, contando com o patrocínio da humanidade como um todo (representada, digamos, pela Assembleia Geral das Nações Unidas), e ainda assim empregar apenas um único cientista para fazer o trabalho.[535]

Existe uma razão para iniciarmos a colaboração o quanto antes: precisamente, tirar vantagem do véu da ignorância que omite qualquer informação específica a respeito de qual projeto individual obterá primeiro uma superinteligência. Quanto mais nos aproximamos da linha de chegada, menor será a incerteza em relação às chances relativas dos projetos que competem entre si e, consequentemente, mais difícil poderia ser convencer o projeto líder de abrir mão de seus interesses próprios para se juntar a um projeto colaborativo que distribuiria os ganhos para toda a humanidade. Por outro lado, também parece complicado estabelecer uma colaboração formal de âmbito global antes que a perspectiva de uma superinteligência seja mais amplamente reconhecida do que é atualmente e num momento anterior ao surgimento de um caminho claro que leve à criação de uma

superinteligência de máquina. Além do mais, na medida em que a colaboração promoveria avanços no decorrer desse caminho, ela poderia ser até mesmo contraproducente em termos de segurança, conforme discutimos anteriormente.

Assim, a forma ideal de colaboração para os dias atuais possivelmente não demandaria acordos formais específicos em seu início e não apressaria o desenvolvimento da inteligência de máquina. Uma proposta que se encaixa nesses critérios é promover uma norma moral apropriada, expressando nosso comprometimento com a ideia de que a superinteligência deveria ser voltada ao bem comum. Tal norma poderia ser formulada da seguinte maneira:

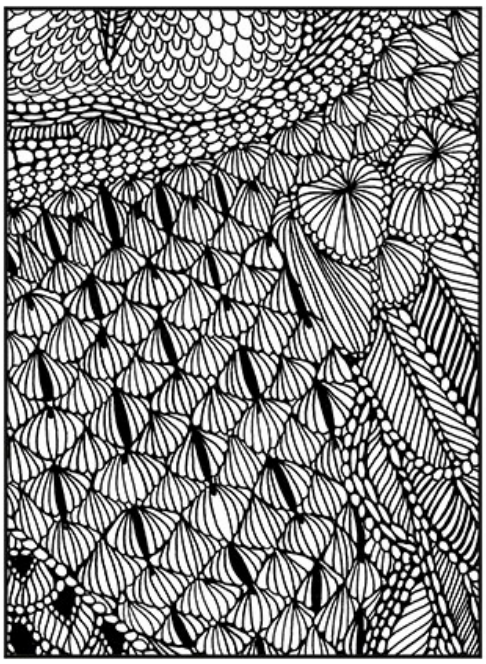
> **O princípio do bem comum**
> Uma superinteligência deverá ser desenvolvida apenas para o benefício de toda a humanidade e a serviço de ideais éticos amplamente compartilhados por todos.[536]

Estabelecer, em seus estágios iniciais, que o imenso potencial de uma superinteligência pertence à toda humanidade forneceria mais tempo para a internalização dessa norma.

O princípio do bem comum não exclui incentivos comerciais para indivíduos e empresas ativas em áreas relacionadas. Por exemplo, uma empresa pode satisfazer o apelo por uma divisão universal dos benefícios de uma superinteligência adotando uma "cláusula de lucro inesperado" como garantia de que todo o lucro obtido até um teto bastante elevado (algo em torno de, digamos, 1 trilhão de dólares anuais) pudesse ser distribuído normalmente entre os acionistas dessa empresa ou a quem mais fosse de direito, e que apenas os lucros que excedessem esse patamar seriam distribuídos entre toda a humanidade de forma igualitária (ou de acordo com outros critérios morais universais). A adoção dessa cláusula de lucro inesperado deveria ser praticamente isenta de custos, sendo extremamente improvável que uma empresa qualquer viesse a exceder esse patamar de lucro estratosférico (e esses cenários de baixa probabilidade normalmente não influenciariam as decisões dos investidores e gerentes da empresa). Contudo, a difusão de sua adoção daria à humanidade uma garantia valiosa (contanto que pudéssemos confiar nesse compromisso) de que, caso alguma empresa ganhasse na loteria da explosão de inteligência, todos receberiam uma parcela da maioria dos benefícios. A mesma ideia poderia ser aplicada a outras entidades além de empresas. Por exemplo, as nações poderiam firmar um acordo garantindo que, caso o PIB de qualquer uma delas exceda uma fração muito elevada (de, digamos, 90%) do PIB mundial, esse excedente seria distribuído igualmente entre todas as outras nações.[537]

O princípio do bem comum (e instanciações particulares, como cláusulas de lucro inesperado) poderia ser adotado inicialmente como um compromisso moral voluntário firmado por indivíduos e organizações responsáveis que sejam ativos em áreas do conhecimento ligadas à inteligência de máquina. Mais tarde, ele poderia ser endossado por um conjunto maior de entidades e transformado em lei ou tratado. Uma formulação vaga, como a fornecida aqui, poderia servir como um

ponto de partida, porém, em algum momento, ela deveria ser aprimorada e convertida em um conjunto de requisitos específicos e verificáveis.



15

CAPÍTULO 15

A HORA DECISIVA

Nos encontramos em um emaranhado de complexidade estratégica, cercados por uma densa névoa de incertezas. Embora muitas considerações tenham sido feitas, seus detalhes e inter-relacionamentos permanecem confusos e incertos — e podem haver outros fatores sobre os quais sequer pensamos a respeito. O que podemos fazer diante dessa difícil situação?

Filosofia com prazo

Um colega gosta de salientar que uma medalha Fields (a mais alta honraria na área da matemática) indica duas coisas em relação ao premiado: que ele tem capacidade de realizar um feito realmente importante e que ele não o realizou. Apesar de ser cruel, essa observação tem uma ponta de verdade.

Pense em uma "descoberta" como um ato que move a chegada da informação de um ponto posterior no tempo até um ponto anterior no tempo. O valor da descoberta não é igual ao valor da informação encontrada, mas equivale ao valor de se ter obtido a informação mais cedo do que, em contrário, poderia ter ocorrido. Um cientista ou um matemático pode mostrar grande habilidade ao ser o primeiro a encontrar uma solução que tenha escapado a muitos outros; ainda assim, se o problema pudesse ser facilmente resolvido de qualquer maneira, seu trabalho não teria beneficiado tanto o mundo. *Há* casos nos quais o fato de se possuir uma solução apenas um pouco mais cedo tem imenso valor, mas isso é mais plausível quando a solução é colocada em uso imediatamente, com algum fim prático ou servindo como base para trabalhos teóricos adicionais. E, nesse último caso, em que uma solução é usada imediatamente apenas no sentido de servir como um bloco de construção para outras teorizações, há grande valor na obtenção de uma solução um pouco mais cedo somente se o trabalho adicional que se possibilitou for tão importante quanto urgente.[538]

Então, a questão não é se o resultado descoberto pelo ganhador da medalha Fields é, por si só, "importante" (instrumentalmente ou em prol do próprio conhecimento). Em vez disso, deveríamos nos perguntar se foi importante que o medalhista tenha possibilitado a publicação do resultado com antecedência. O valor desse transporte temporal deveria ser comparado ao valor daquilo que uma mente matemática de excelência poderia ter gerado caso estivesse trabalhando em alguma outra questão. Pelo menos em alguns casos, a medalha Fields pode indicar uma vida inteira consumida na resolução do problema errado — por exemplo, um problema cujo fascínio esteja apenas no fato de ser reconhecidamente de difícil solução.

Farpas semelhantes poderiam ser dirigidas a outras áreas, tais como a filosofia acadêmica. A filosofia abrange alguns problemas relevantes para a mitigação do risco existencial — nós encontramos muitos neste livro. Porém, também há subcampos dentro da filosofia que aparentemente não têm ligação com o risco

existencial ou, de fato, quaisquer outras preocupações práticas. Assim como a matemática pura, alguns dos problemas que a filosofia estuda podem ser considerados intrinsecamente importantes, pois os humanos têm razão para se preocupar com eles independentemente de qualquer aplicação prática. A natureza fundamental da realidade, por exemplo, vale a pena ser conhecida por si só. O mundo seria possivelmente menos glorioso se ninguém estudasse metafísica, cosmologia ou a teoria das cordas. Entretanto, a perspectiva nascente de uma explosão de inteligência lança uma nova luz sobre essa antiga busca pela sabedoria.

Existe atualmente um cenário que sugere a maximização do progresso filosófico por meio de um trajeto indireto em vez de um filosofia imediata. Uma das muitas tarefas nas quais uma superinteligência (ou apenas uma inteligência humana moderadamente aperfeiçoada) seria capaz de superar o grupo de pensadores atuais seria a de responder perguntas fundamentais de ciência e filosofia. Essa reflexão sugere uma estratégia de adiamento de gratificação. Por enquanto, poderíamos adiar um pouco o trabalho em torno de algumas das perguntas eternas, delegando essa tarefa a sucessores que esperamos ser mais competentes — para que possamos nos concentrar num desafio mais urgente: aumentar a chance de que realmente teremos sucessores competentes. Isso seria filosofia de alto impacto e matemática de alto impacto.[539]

O que deve ser feito?

A intenção, portanto, é nos concentrar em problemas que não são apenas importantes, mas urgentes, pois suas soluções seriam necessárias antes da explosão de inteligência. Também deveríamos tomar cuidado para não trabalhar com problemas que apresentem valor negativo (uma vez que resolvê-los é algo perigoso). Alguns problemas técnicos na área da inteligência artificial, por exemplo, podem ter valor negativo, uma vez que suas soluções acelerariam o desenvolvimento da superinteligência de máquina sem, ao mesmo tempo, agilizar o desenvolvimento de métodos de controle que poderiam tornar a revolução da inteligência de máquina benéfica, possibilitando que sobrevivêssemos a ela.

Pode ser difícil identificar problemas que são tanto urgentes quanto importantes e cujo valor positivo seja uma certeza. A incerteza estratégica em torno da mitigação do risco existencial indica que devemos nos preocupar com o fato de que mesmo intervenções bem-intencionadas podem acabar se revelando não apenas improdutivas, mas também contraproducentes. Para limitar o risco de que algo ativamente nocivo ou moralmente incorreto seja realizado, devemos dar preferência ao trabalho com problemas que aparentemente tenham *valor positivo robusto* (por exemplo, aqueles cujas soluções forneceriam uma contribuição positiva em uma grande variedade de situações) e empregar meios que fossem fortemente justificáveis (por exemplo, aceitáveis de acordo com uma grande variedade de visões morais).

Há mais um desiderato a ser considerado na seleção de problemas que devemos priorizar. Queremos trabalhar com problemas que sejam *elásticos* em resposta aos

nossos esforços para resolvê-los. Problemas altamente elásticos são aqueles que podem ser resolvidos muito mais rapidamente, ou em um âmbito muito maior, com uma unidade extra de esforço. Estimular a bondade no mundo é um problema urgente e importante — um problema que, mais do que nunca, parece ter um valor positivo bem sólido: mas, na ausência de uma descoberta que indique como fazer isso, este se torna provavelmente um problema de elasticidade bem baixa. Da mesma forma, alcançar a paz mundial seria algo altamente desejável; porém, considerando os numerosos esforços já direcionados para esse problema e os formidáveis obstáculos a uma solução rápida, parece improvável que as contribuições de alguns indivíduos a mais poderiam fazer uma grande diferença.

Com a finalidade de reduzir os riscos da revolução de inteligência de máquina, serão propostos dois objetivos que parecem sintetizar da melhor forma possível todos esses desideratos: análise estratégica e construção de capacidades. Podemos nos encontrar relativamente confiantes em relação ao sinal desses parâmetros — quanto mais entendimento estratégico e mais construção de capacidades, melhor. Além disso, esses parâmetros são elásticos: um pequeno investimento extra pode fazer uma diferença relativamente grande. Obter entendimento e capacidades também é urgente, pois ganhos iniciais nesses parâmetros podem se acumular, tornado mais efetivos os esforços subsequentes. Além desses dois objetivos amplos, apontaremos alguns outros alvos que potencialmente merecem iniciativas.

Buscando a iluminação estratégica

Para confrontar um cenário de perplexidade e incerteza, a análise estratégica se destaca com um valor esperado particularmente alto.[540] Iluminar nossa situação estratégica nos ajudaria a escolher intervenções subsequentes de modo mais efetivo. A análise estratégica é especialmente necessária quando estamos totalmente incertos não apenas sobre alguns detalhes relacionados a alguma questão periférica, mas também em relação às qualidades cardinais das questões centrais. Estamos totalmente incertos a respeito de diversos parâmetros essenciais, até mesmo em relação ao seu *sinal* — ou seja, não sabemos em qual direção as mudanças seriam desejáveis e em qual se apresentariam indesejáveis. A nossa ignorância pode não ser irremediável. Esse tema tem sido pouco estudado e insights estratégicos brilhantes podem estar à espera, a poucos metros da superfície, de serem desenterrados.

Com “análise estratégica”, nos referimos à busca por *considerações cruciais*: ideias ou argumentos que tenham o potencial de modificar nossas visões, não apenas a respeito da estrutura detalhada de implementação, mas também em relação à topologia geral de conveniência.[541] Mesmo uma única consideração crucial perdida poderia viciar os nossos esforços mais corajosos ou torná-los ativamente perigosos, como os de um soldado que estivesse lutando no lado errado. A busca por considerações cruciais (que deve explorar questões normativas e descritivas) demandará frequentemente o cruzamento das fronteiras entre diferentes disciplinas acadêmicas e outros campos do conhecimento. Como não há

nenhuma metodologia estabelecida para realizar esse tipo de pesquisa, precisaremos de um raciocínio difícil e original.

Construindo uma boa capacidade

Outra atividade de alto valor, que compartilha com a análise estratégica a mesma propriedade robusta de ser benéfica em uma ampla gama de cenários, é o desenvolvimento de uma base de apoio bem constituída e que leve o futuro a sério. Tal base de apoio poderia fornecer imediatamente recursos para pesquisa e análise. Se e quando outras prioridades se tornassem visíveis, esses recursos poderiam ser redirecionados de acordo com essas prioridades. Uma base de apoio é, portanto, uma capacidade de propósito geral cujo uso pode ser guiado por novos insights conforme forem surgindo.

Um recurso valioso seria uma rede de doadores formada por indivíduos dedicados à filantropia racional, informados a respeito de riscos existenciais e atentos às formas de mitigá-los. Seria preferível que os primeiros financiadores fossem astutos e altruístas, pois eles teriam oportunidades de moldar a cultura da área antes que os interesses venais usuais assumissem suas posições e se enraizassem. Durante essas manobras iniciais, deveríamos, portanto, nos concentrar em recrutar os tipos certos de pessoas para a área. Valeria a pena renunciar a alguns avanços técnicos no curto prazo para preencher essas posições com indivíduos que genuinamente se importem com a segurança e que tenham uma vocação para a busca da verdade (e que provavelmente atrairão mais pessoas como eles).

Uma variável importante é a qualidade da "epistemologia social" do campo da IA e de seus principais projetos. Descobrir considerações cruciais é valioso, mas apenas se essas descobertas afetarem as ações. Nunca devemos deixar de atribuir o devido valor a esse ponto. Imagine um projeto que tenha investido milhões de dólares e muitos anos de labuta no desenvolvimento de um protótipo de IA e que, depois de ultrapassados muitos desafios técnicos, o sistema estivesse começando finalmente a mostrar um progresso real. Há a possibilidade de que o projeto poderia se tornar algo útil e rentável apenas com um pouco mais de trabalho. Então uma consideração crucial é descoberta, indicando que uma abordagem completamente diferente seria um pouco mais segura. O projeto se autodestruiria como um samurai desonrado, abandonando seu design inseguro e todo o progresso já alcançado? Ou reagiria como um polvo apreensivo, soltando uma nuvem preta de ceticismo na esperança de escapar do ataque? Um projeto que escolhesse de forma confiável a opção do samurai em um dilema desse tipo seria bem mais favorável.[542] Entretanto, construir processos e instituições que estejam dispostos a cometer um haraquiri com base em alegações incertas e razões especulativas não é algo fácil. Outra dimensão da epistemologia social é o gerenciamento de informações sensíveis, em particular a habilidade de evitar o vazamento de informações que deveriam ser mantidas em segredo. (A continência de informações pode ser especialmente desafiadora para pesquisadores acadêmicos, acostumados a pregar o resultado de suas pesquisas em cada poste de luz e árvores que estejam disponíveis.)

Medidas específicas

Além dos objetivos gerais de iluminação estratégica e boa capacidade, alguns objetivos mais específicos também poderiam apresentar oportunidades de ação com um bom custo-benefício.

Um deles é o progresso em relação aos desafios técnicos de segurança da inteligência de máquina. Na busca por esse objetivo, devemos tomar o cuidado de gerenciar perigos de informação. Alguns dos avanços que seriam úteis na resolução do problema do controle também serviriam para resolver o problema da competência. Um avanço que queimasse o fusível da IA poderia facilmente ter um valor líquido negativo.

Outro objetivo específico é a promoção de "melhores práticas" entre os pesquisadores de IA. Qualquer progresso feito em relação ao problema do controle precisa ser disseminado. Algumas formas de experimentação computacional, particularmente aquelas que envolvem o uso de autoaperfeiçoamento recursivo forte, também poderiam demandar o uso do controle de capacidade para atenuar o risco de uma partida acidental. Embora a implementação real de métodos de segurança não seja algo tão relevante hoje em dia, ela se tornará gradativamente relevante na medida em que o estado da arte avançar. E não é demasiado cedo para convocar os profissionais a declararem um *compromisso com a segurança*, incluindo o endossamento do princípio do bem comum e a promessa de intensificar a segurança se e quando a perspectiva de uma superinteligência de máquina começar a parecer mais iminente. Palavras bondosas não são suficientes e não tornarão, por si só, segura uma tecnologia perigosa: mas o pensamento poderia gradualmente seguir na direção das palavras.

Outras oportunidades de modificar algum parâmetro crucial também surgiriam ocasionalmente, por exemplo, para atenuar algum outro risco existencial, promover aprimoramentos cognitivos biológicos e melhorias da nossa sabedoria coletiva, ou até mesmo para modificar a política mundial em direção a algo mais harmonioso.

Que o melhor da natureza humana se coloque, por favor, de pé

Diante do prospecto de uma explosão da inteligência, nós, humanos, somos como crianças pequenas brincando com uma bomba. Tamanho é o descompasso entre o poder do nosso brinquedo e a imaturidade da nossa conduta. A superinteligência é um desafio para o qual não estamos preparados atualmente e assim continuaremos por um longo tempo. Sabemos pouco a respeito do momento em que a detonação ocorrerá, embora seja possível ouvir um fraco tique-taque quando aproximamos o dispositivo dos nossos ouvidos.

No caso de uma criança com uma bomba não detonada em suas mãos, uma atitude sensata seria colocá-la no chão com muito cuidado, sair rapidamente da sala e contatar o adulto mais próximo. Porém, não temos aqui apenas uma criança, mas muitas, cada uma com acesso a um mecanismo detonador independente. As chances de que *todos nós* teremos o bom senso de depositar o material perigoso

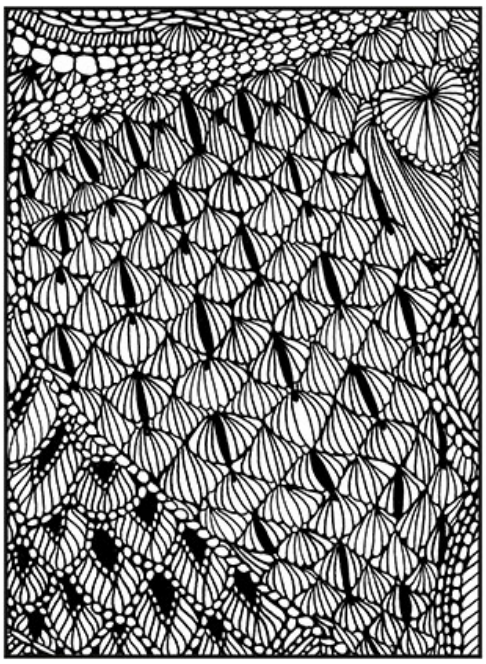
no chão parecem quase nulas. Algum pequeno idiota inevitavelmente apertará o botão de ignição só para ver o que acontece.

Também não seria possível obter segurança fugindo dessa situação, pois a detonação de uma explosão de inteligência colocaria o firmamento inteiro abaixo. E, por fim, não há nenhum adulto à vista.

Nesse caso, nenhum sentimento eufórico de surpresa faria sentido. Consternação e medo seriam sentimentos muito mais pertinentes; mas a atitude mais apropriada poderia ser uma determinação árdua de atingir nossa máxima competência, como se estivéssemos nos preparando para uma prova muito difícil que poderia tanto realizar nossos sonhos como destruí-los.

Essa não é uma prescrição de fanatismo. A explosão de inteligência pode estar ainda muitas décadas à frente. Além disso, em certa medida, nosso desafio será nos agarrar à nossa humanidade: manter nossos fundamentos, nosso bom senso e uma decência bem-humorada, mesmo diante do enfrentamento desse problema que é o mais desumano e artificial dos problemas. Precisamos juntar todas as nossas capacidades humanas para alcançar sua solução.

No entanto, não percamos de vista aquilo que tem importância global. Através da neblina das trivialidades cotidianas, podemos entrever — ainda que vagamente — a tarefa essencial da nossa era. Neste livro, tentamos identificar um pouco melhor as características daquilo que, de outra forma, ainda se mostra como uma visão relativamente amorfa e negativamente definida — uma visão que apresenta como a nossa principal prioridade moral (pelos menos em uma perspectiva impessoal e secular) a redução do risco existencial e a realização de uma trajetória da civilização que conduziria a um uso solidário e exultante da herança cósmica da humanidade.



SUPERINTELIGÊNCIA

POSFÁCIO

Desde a publicação da primeira edição deste livro, ocorreram algumas mudanças de atitude. Ficou mais fácil tratar a superinteligência como um tópico importante — levar a sério a visão de que a transição para a inteligência de máquina pode acontecer neste século, que tal transição pode estar entre os eventos mais importantes da história da humanidade, que ela pode vir acompanhada de uma certa dose de risco existencial, embora também possa ser extremamente positiva, e que seria sensato adiantar um pouco o trabalho de buscar entender se há algo que deveríamos estar fazendo para aumentar as chances de um resultado favorável. Sabemos que ainda existe a imagem de um Exterminador do Futuro zombando de praticamente qualquer tentativa jornalística dedicada ao tema. Mas, afastando-nos da cacofonia popular, atualmente já é possível — se apontarmos nossos ouvidos na direção certa — escutar o burburinho de uma conversa mais madura.

Avanços técnicos em aprendizado de máquina têm ocorrido mais rapidamente do que muitos esperavam. Uma vasta fronteira de ideias por explorar foi possibilitada por desenvolvimentos recentes: máquinas neurais de Turing, aprendizagem por reforço profunda, otimização bayesiana de hiperparâmetros, *grids* de redes LSTM, redes de memória, autocodificadores variacionais, incorporação de vetores em nível de sentença, redes generativas adversariais, modelos generativos baseados em atenção, diversas abordagens em programação probabilística — para elencar apenas algumas das fontes do atual burburinho. A aprendizagem profunda tem sido o foco de grande parte do entusiasmo. Métodos de aprendizagem profunda — essencialmente redes neurais com múltiplas camadas — têm começado, graças a uma combinação de computadores mais rápidos, bases de dados maiores e refinamentos algorítmicos, a se aproximar (e em alguns casos a exceder) o desempenho humano em diversas tarefas de percepção, incluindo reconhecimento de escrita, reconhecimento e descrição de imagens, reconhecimento de voz e reconhecimento facial. Métodos de aprendizagem profunda também vêm alcançando resultados expressivos em tradução de linguagem natural e em alguns problemas de análise de dados científicos. A capacidade que está por trás desses desempenhos — algoritmos genéricos capazes de aprender representações distribuídas abstratas a partir de dados brutos de sensores sem que um humano precise programar manualmente características relevantes ou dispensando um conhecimento específico do domínio — pode se tornar um alicerce importante na construção de capacidades mais complexas.

Muitas das capacidades existentes alcançaram o limiar de utilidade. Isso significa que há uma recompensa imediata para novos avanços, já que eles se traduzirão diretamente em melhorias para produtos comercialmente valiosos. Se de início você tiver um sistema de reconhecimento de voz tão impreciso que não apresente nenhuma utilidade e posteriormente faça pequenas melhorias nele, bem, você continuará com um sistema inútil. Mas se logo no início você tiver um sistema que seja bom o suficiente para ser amplamente utilizado, uma melhoria de 1% pode

valer 1 bilhão de dólares. A emocionante sensação de que o aprendizado de máquina se encontra em movimento, com muitos caminhos ainda por explorar, atualmente também tem sido incrivelmente galvanizada por incentivos comerciais e vem atraindo muito dinheiro e talentos para essa área.

O quão longe poderá ir essa atual onda de empolgação, ninguém sabe dizer. Não faz parte da argumentação deste livro afirmar que uma explosão de inteligência esteja iminente ou que o mundo tenha subestimado o ritmo do progresso. É perfeitamente factível que os avanços passíveis de se obter através da intensificação e do aprimoramento das abordagens atuais alcancem, em breve, seus limites. Certamente, novas ideias e descobertas serão necessárias para que possamos percorrer o restante do caminho — ideias que podem ou não surgir a tempo. No entanto, eu ficaria um pouco surpreso se testemunhasse a ocorrência de um novo "inverno da IA" tão severo quanto aqueles enfrentados no passado. Parece mais provável que haja investimentos contínuos e a conquista de uma maior reputação desse ponto em diante, uma vez que já alcançamos o suficiente para estabelecer a inteligência artificial tanto como uma área promissora de estudos teóricos quanto como uma disciplina útil da área de engenharia. (Esse prognóstico é, obviamente, inteiramente consistente com o estouro de uma ou outra bolha de investimento e com a incapacidade, por parte da maioria dos projetos individuais, de fazer jus às expectativas de seus proponentes.)

Também houve algum progresso no estudo das maneiras de aumentar as chances de um resultado favorável. No campo teórico, foram introduzidas algumas ideias novas (por exemplo, os "agentes direcionados à aprovação" de Paul Christiano) que merecem maiores estudos. No campo organizacional, a situação também melhorou um pouco — essa melhoria é substancial em termos relativos, mas está se iniciando a partir de um nível anterior bastante baixo. O reconhecimento do problema do controle enquanto um tópico de investigação legítimo tem se espalhado por uma fração menos minúscula das comunidades científicas relevantes, e há um reconhecimento mais amplo de que os impactos futuros da inteligência de máquina merecem uma atenção mais sistemática.

Os avanços em relação à criação das áreas de estudos sobre segurança e impactos da IA, apesar de notáveis por terem ocorrido em um período de tempo tão curto, não devem ser superestimados. De fato, há mais financiamento na área, mas esse financiamento ainda é duas ou três ordens de magnitude menor do que aquele direcionado para o desenvolvimento puro e simples de máquinas mais inteligentes. Sim, há um maior interesse em se pensar a respeito das consequências dos avanços em inteligência de máquina, mas muito desse interesse acaba se concentrando em preocupações de curto prazo, como máquinas autônomas letais, impactos da automação no mercado de trabalho, cibercriminalidade, privacidade ou carros autônomos. Não é despropositado que algumas pessoas estejam pensando a respeito desses assuntos, mas, na maioria dos casos, elas se preocupam com questões bem distintas daquelas que surgem nos casos de uma IA de nível humano ou de uma superinteligência.

E, sim, tem surgido na comunidade de pesquisadores de IA um certo grau de legitimidade em torno de preocupações de longo prazo. Mas essa aceitação é ainda

muito parcial — e, provavelmente, fraca. Alguns pesquisadores de IA estão preocupados com o fato de que a opinião pública possa estar começando a sair do controle. Os filmes sem sentido no estilo *Exterminador do Futuro* estão surtindo efeito. Não deve ser muito divertido ter calúnias despejadas sobre sua disciplina acadêmica, sua comunidade profissional, seu trabalho de toda uma vida. Existe a possibilidade de que, em resposta a esse alarmismo equivocado em torno de exércitos de robôs malvados, a comunidade de IA se coloque fortemente numa postura de depreciação de quaisquer preocupações reais a respeito da inteligência de máquina avançada. Essa postura pode vir a ser útil, contanto que se direcione especificamente ao alarmismo criado pela mídia e ajude a diminuí-lo um pouco. Mas há a preocupação de que isso cause algum dano colateral. Se os pesquisadores de IA estabelecerem uma norma segundo a qual seria pouco prudente falar a respeito da superinteligência ou questionar seus possíveis riscos, com receio de "fornecer munição" aos críticos, alarmistas, fanáticos ou aspirantes a fiscais, nesse caso, os ganhos recentes em legitimidade poderiam ser rapidamente revertidos. Poderíamos, então, entrar em um "inverno da segurança da IA", um clima no qual seria mais difícil realizar o tipo de trabalho proposto neste livro. É importante impedir que tal dinâmica antagonista ganhe força. O melhor caminho para o desenvolvimento de uma superinteligência benéfica é aquele no qual desenvolvedores de IA e pesquisadores de segurança de IA estejam do mesmo lado — onde eles até possam ser, em grande medida, as mesmas pessoas. Sendo assim, eu conclamo todos os lados a exercerem a paciência e a moderação, a manter a mente aberta e a se empenhar em travar um diálogo franco e colaborar sempre que possível.

Abrirei mão de responder aqui a todos os comentários que foram feitos sobre este livro desde sua publicação. Farei apenas uma pequena observação — dirigida aos leitores desta nova edição e, principalmente, àqueles cujas vidas se tornaram tão atarefadas que eles de fato pararam de *ler* os livros que adquirem, salvo, talvez, por uma olhada de relance no índice e nas primeiras e últimas páginas: faço essa observação para salientar que há muitos fatores, além da importância intrínseca de uma questão, que podem influenciar no número de páginas de um livro. Sendo assim, não é possível inferir minhas crenças com base numa simples contagem de páginas. Por exemplo, eu me concentro mais nos riscos do que nos benefícios potenciais. Isso não significa que eu não considere esses últimos enormes; apenas acredito que, neste momento da história, quando podemos nos satisfazer com uma vaga noção de que existem possíveis benefícios (astronômicos) caso a transição para inteligência de máquina transcorra bem, parece ser mais urgente desenvolver um conhecimento detalhado e preciso de tudo aquilo que especificamente poderia dar errado — para que possamos ter certeza de que isso poderá ser evitado. Do mesmo modo, o fato de eu ter ocupado inúmeras páginas analisando cenários nos quais uma única IA superinteligente seria poderosa o suficiente para moldar o futuro de acordo com suas preferências, não significa que eu descarte a possibilidade de cenários multipolares (ver, por exemplo, o capítulo 11). Também discorro longamente a respeito das possíveis dificuldades relacionadas ao problema do controle e sobre as maneiras pelas quais soluções superficialmente plausíveis

poderiam falhar, mas o problema do controle poderia, igualmente, se revelar de fácil solução.

Gostaria novamente de agradecer a todos os que me auxiliaram na criação deste livro ou que contribuíram para a sua receptividade — e a todos aqueles mundo afora que tentam atuar construtivamente nessa nossa difícil e estranha situação humana.

Nick Bostrom, *novembro de 2015*

SUPERINTELIGÊNCIA

GLOSSÁRIO PARCIAL

Abordagem de seleção evolutiva para a inserção de valores: Uma abordagem para o problema de inserção de valores que busca obter uma IA com valores desejáveis através de um processo de seleção iterativo, semelhante ao processo evolutivo de seleção que produziu os seres humanos.

Acelerador de desenvolvimento macroestrutural: Uma alavanca imaginária (utilizada em experimentos mentais) que alteraria o ritmo de desenvolvimento de características macroestruturais (tais como tecnologia e dinâmica geopolítica), enquanto se manteria inalterado o ritmo com o qual eventos humanos ocorridos em um nível micro se desenrolariam.

Acoplamento de tecnologias: Uma relação cronológica previsível entre duas tecnologias, de tal modo que o desenvolvimento de uma tecnologia tenha uma forte tendência de levar ao desenvolvimento de outra, tanto como um precursor necessário ou como uma inevitável e irresistível etapa subsequente. Por exemplo, a emulação completa do cérebro está, plausivelmente, acoplada à IA neuromórfica, porque uma versão mais primitiva da tecnologia que seria necessária para uma emulação completa do cérebro já poderia possibilitar a criação de uma IA inspirada no cérebro (e haveria fortes incentivos para que essa tecnologia fosse usada para esse fim).

Acréscimo associativo de valores: Uma abordagem para a inserção de valores na qual um mecanismo é especificado de modo a fazer com que a IA adote, ao longo do tempo, novos valores finais de uma maneira que seja dependente de suas experiências. (Inspirada na maneira pela qual os humanos, aparentemente, adquirem seus valores.)

Ampliação: Uma abordagem para a obtenção de uma superinteligência com motivações desejáveis que consiste em iniciar com um sistema que já possua motivações apropriadas (por exemplo, um ser humano) e torná-lo mais inteligente, em vez de tentar criar um sistema de objetivos do zero.

Aprendizado de valores: Uma abordagem para o problema de inserção de valores na qual a IA aprende os valores que os humanos queiram que ela siga.

Aprendizado por reforço aplicado ao problema do controle: Uma abordagem na qual a IA aprende a maximizar alguma noção de recompensa cumulativa (onde o sinal de recompensa é especificado ou administrado por humanos de algum modo que objetive induzir a IA a se comportar da maneira desejada).

Aprimoramento cognitivo: Melhorias nas habilidades intelectuais de um sistema.

Autoaperfeiçoamento recursivo: O processo pelo qual uma IA (talvez uma IA embrionária) aumenta iterativamente sua própria inteligência — utilizando esse aumento de inteligência para atribuir cada vez mais poder de otimização ao aumento de sua própria cognição.

Captura antrópica: Um fenômeno hipotético no qual uma IA pensa que poderia estar em uma simulação e, por causa disso, ela tentaria se comportar de maneiras que seriam recompensadas por seus simuladores.

Confinamento: Um método de controle que consiste em projetar o ambiente de uma IA de modo a reduzir sua capacidade de interação com o mundo exterior, por exemplo, executando a IA numa realidade virtual simulada, cuidadosamente isolada e com acesso apenas a um canal restrito de comunicação com seus supervisores humanos.

Conjectura de realização tecnológica: Se os esforços para o desenvolvimento científico e tecnológico não forem efetivamente encerrados, então todas as capacidades básicas importantes que poderiam ser obtidas por meio de alguma tecnologia possível o serão.

Consideração crucial: Uma ideia ou argumento que poderia plausivelmente revelar a necessidade não apenas de pequenos ajustes de curso em nossos esforços práticos, mas de uma grande mudança de direção ou prioridade, fazendo com que, por exemplo, intervenções importantes que eram consideradas positivas passem a ser consideradas negativas, ou vice-versa.

Crimes da mente: O abuso de processos computacionais moralmente relevantes (em simulações, ou dentro de uma IA, ou criados em substratos de máquina por razões instrumentais).

Domesticidade: Uma abordagem para o problema do controle na qual o sistema de motivação da IA é projetado de modo que ela tenha ambições muito modestas, limitando a gama de aspectos nos quais ela tenha vontade de interferir.

Emulação completa do cérebro: Inteligência de máquina criada através da cópia da estrutura computacional do cérebro humano.

Especificação direta: Uma abordagem para o problema do controle na qual os programadores descobrem o que os humanos valorizam e, então, programam a IA explicitamente com os valores ou regras correspondentes.

Estepe motivacional: Uma abordagem para o problema de inserção de valores na qual uma IA embrionária recebe, primeiramente, objetivos simples que são, posteriormente, substituídos por outros objetivos mais complexos (alinhados a um valor), uma vez que a IA tenha desenvolvido recursos representacionais suficientemente sofisticados.

Excedente de hardware: Ocorre com o surgimento de processos algorítmicos mais sofisticados que são capazes de utilizar de maneira mais eficiente o hardware já em existência, sem que seja necessário um aumento físico da capacidade desse hardware.
Explosão de inteligência: Um evento hipotético no qual uma IA, partindo de um nível "relativamente modesto", se aperfeiçoa rapidamente, atingindo um nível de inteligência radicalmente sobre-humano (um processo no qual, supostamente, será empregado o autoaperfeiçoamento recursivo).
Gênio: Uma IA que executa um comando de alto nível e, em seguida, fica à espera do próximo comando.
Hardware de nível humano: Um hardware que se equivale ao cérebro humano em habilidade de processamento de informação.
Herança cósmica da humanidade: O reservatório de recursos físicos presentes no universo que é acessível a uma civilização tecnologicamente madura originária da Terra (começando a partir da Terra atual).
IA de nível humano: Uma IA que tenha desempenho intelectual equivalente ao de um ser humano típico nos principais campos do conhecimento. (Esse termo é significativamente ambíguo.)
IA de retidão moral (**RM**): Uma IA que busca fazer o que é moralmente correto.
IA embrionária: Uma IA com capacidades iniciais modestas que consegue se tornar uma IA expressiva através do aprimoramento de sua própria arquitetura.
IA ferramenta: Uma IA que "não se parece com um agente", mas sim com uma versão mais flexível e capacitada de um software contemporâneo. Especificamente, ela não é direcionada a objetivos.
Inibição: Um método de controle que consiste em limitar as capacidades internas da IA, tal como restringir seu acesso à informação ou criar limitações para suas faculdades cognitivas.
Instanciação perversa: Uma maneira eficiente de satisfazer um objetivo de uma IA que viola as intenções dos programadores responsáveis pela definição dos critérios desse objetivo (por exemplo, alcançar o objetivo de "fazer os humanos sorrirem" paralisando os músculos faciais humanos em sorrisos perpétuos.)
Métodos de controle de capacidade: Estratégias que limitam o que a IA pode fazer como forma de evitar resultados indesejáveis.
Métodos de incentivo: Estratégias de controle de uma IA que consistem em configurar o ambiente dessa IA de tal modo que ela possua razões instrumentais, ainda que se seus objetivos finais não estejam alinhados com os valores humanos.
Modulação de emulações: Inicia-se a partir de emulações cerebrais com motivações aproximadamente humanas que posteriormente são modificadas por meio de um análogo digital de uma droga ou por outros meios.
Normatividade indireta: Uma abordagem para o problema da seleção de valores na qual, em vez de nós mesmos determinarmos diretamente quais valores uma IA deve promover, são especificados um critério ou método que a IA possa seguir, utilizando-se de seus próprios recursos intelectuais para determinar o conteúdo concreto de um padrão normativo definido apenas implicitamente.
O princípio do bem comum: Uma superinteligência deve ser desenvolvida apenas em benefício da humanidade e a serviço de ideais morais amplamente compartilhadas.
Oráculo: Uma IA que somente responde a perguntas.
Pacote de incentivos: Provisões nos objetivos dados a uma IA que alocam recompensas adicionais ou influência para aqueles que auxiliaram em seu desenvolvimento.
Partida: A transição de uma condição na qual existam apenas inteligências de máquina de nível humano para uma na qual existam superinteligências radicais. Ela é geralmente caraterizada por sua velocidade: uma "partida lenta", dura décadas ou séculos; uma "partida moderada", dura meses ou anos; e uma "partida rápida", dura alguns dias ou menos.
Patamar de sustentabilidade de um *singleton* prudente: Um conjunto de capacidades excede o patamar de um singleton se e apenas se um sistema paciente e capaz de evitar riscos existenciais com esse conjunto de capacidades poderia, caso não enfrentasse nenhuma oposição ou competição inteligente, ser capaz de colonizar e reprojetar uma grande parte do universo acessível.
Perspectiva do indivíduo-afetado: A ideia de que deveríamos agir de acordo com os interesses de todas as pessoas que existem atualmente ou que virão a existir independentemente de nossas escolhas (comparar com a "perspectiva impessoal").
Perspectiva impessoal: A ideia de que deveríamos agir em benefício de todos, incluindo aqueles que poderiam vir a existir, com base em nossas escolhas (comparar com a "perspectiva do indivíduo afetado").
Poder de otimização: A quantidade de esforço de projeto, ajustada pela qualidade, que esteja sendo aplicada para aprimorar a inteligência de um sistema.
Primeiro problema de diretor-agente: O conhecido problema enfrentado por uma entidade humana ("o diretor") que designa outra entidade ("o agente") para agir segundo os interesses da primeira (encontrado, por exemplo, na relação entre empregador e empregado).
Princípio da deferência epistêmica: Uma superinteligência futura ocupa uma posição epistemologicamente superior: suas crenças têm (provavelmente, na maioria dos tópicos) mais chance de serem verdadeiras do que

as nossas. Deveríamos, portanto, deferir à opinião da superinteligência sempre que possível.

Princípio do desenvolvimento tecnológico diferencial: Retardar o desenvolvimento de tecnologias nocivas ou perigosas, especialmente aquelas que aumentam o nível de risco existencial e acelerar o desenvolvimento de tecnologias benéficas, especialmente aquelas que reduzam os riscos existenciais associados à natureza ou a outras tecnologias.

Problema de inserção de valores: O problema de como fazer uma IA perseguir como objetivo final os valores que queremos que ele tenha.

Resistência: A dificuldade que um sistema apresenta em ser aprimorado.

Resultado multipolar: Uma condição após a transição para a era da inteligência de máquina na qual existam múltiplos agentes superinteligentes competindo entre si.

Risco de etapa: Um risco associado a uma transição. Nesse caso, a quantidade de risco não é uma função simples do tempo de duração dessa transição. Por exemplo, a travessia de um campo minado não se torna mais segura se for realizada mais rapidamente.

Risco de situação: Um risco associado ao fato de nos encontrarmos em uma determinada situação, de tal modo que a quantidade de risco seja uma função direta do tempo que permanecermos nessa situação. Por exemplo, uma situação na qual não possuímos uma tecnologia de defesa contra o impacto de asteroides apresenta um risco proporcional ao período em que permanecemos nessa situação.

Segundo problema de diretor-agente: O problema de diretor-agente confrontado por um humano “diretor” que deseje desenvolver uma IA superinteligente (“agente”) para que ela trabalhe a seu favor (do humano). Também chamado de “problema do controle”.

Singleton: Uma ordem mundial na qual exista, no nível mais alto de tomadas de decisão, uma única agência (mesmo que internamente essa agência contenha diversas facções e interesses distintos), ou seja, uma condição na qual todos os problemas relevantes de coordenação global estejam praticamente solucionados. Possíveis exemplos incluem uma democracia mundial, uma ditadura mundial absoluta ou uma IA superinteligente poderosa o suficiente para eliminar qualquer rival em potencial.

Soberano: Uma IA que age de maneira autônoma no mundo em busca de objetivos amplos e possivelmente de muito longo prazo.

Software de nível humano: Um software que se equivale a um cérebro humano em termos de eficiência algorítmica na execução de tarefas realizadas por um cérebro humano.

Superinteligência coletiva: Um sistema composto de um grande número de intelectos menores de modo que seu desempenho geral supere em muito qualquer sistema cognitivo atual em diversas áreas gerais de conhecimento.

Superinteligência de qualidade: Um sistema que seja pelo menos tão rápido quanto uma mente humana, mas qualitativamente muitíssimo mais inteligente.

Superinteligência rápida: Um sistema que pode fazer tudo o que um intelecto humano faz, porém muito mais rapidamente.

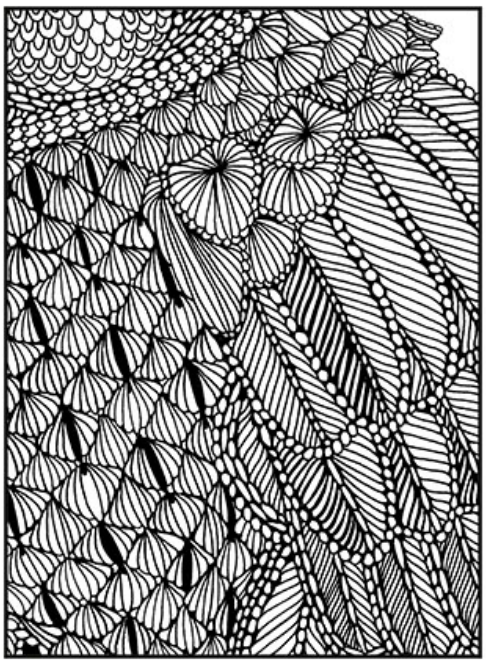
Superinteligência: Qualquer intelecto que exceda em muito a performance cognitiva dos humanos em virtualmente qualquer área do conhecimento.

Tese da convergência instrumental: Podemos identificar “valores instrumentais convergentes”, objetivos secundários que sejam úteis para a obtenção de uma ampla gama de possíveis objetivos finais em uma ampla gama de possíveis ambientes — objetivos secundários que seriam, portanto, mais provavelmente perseguidos por uma classe extensa de agentes inteligentes.

Tese da ortogonalidade: Inteligência e objetivos finais são ortogonais: praticamente qualquer nível de inteligência poderia, em princípio, ser combinado com praticamente qualquer tipo de objetivo final.

Vantagem estratégica decisiva: Superioridade estratégica (tecnológica ou por algum outro meio) suficiente para capacitar um agente a dominar completamente o mundo.

Esse glossário foi incluído apenas para conveniência, embora não esteja nem completo (muitos termos e conceitos diferenciados que foram introduzidos ao longo do livro foram omitidos) nem seja dispositivo (as discussões feitas ao longo do texto principal têm preferência nos casos onde não fica claro como alguma noção deve ser entendida). (Sou grato a Stephanie Zolayvar e Katja Gracie, membros do AI Impacts, Machine Intelligence Research Institute, por terem feito uma versão inicial desse glossário e permitido que eu a modificasse. A versão original pode ser encontrada em: http://aiimpacts.org/ai-risk-terminology/.)



SUPERINTELIGÊNCIA

BIBLIOGRAFIA

ACEMOGLU, Daron. 2003. “Labor- and Capital-Augmenting Technical Change”. *Journal of the European Economic Association* 1 (1): 1-37.

ALBERTSON, D. G. e Thomson, J. N. 1976. “The Pharynx of *Caenorhabditis Elegans*”. *Philosophical Transactions of the Royal Society B: Biological Sciences* 275 (938): 299-325.

ALLEN, Robert C. 2008. “A Review of Gregory Clark’s *A Farewell to Alms*: *A Brief Economic History of the World*”. *Journal of Economic Literature* 46 (4): 946-73.

AMERICAN HORSE COUNCIL. 2005. “National Economic Impact of the US Horse Industry”. Recuperado em jul. 30, 2013. Disponível em www.horsecouncil.org/national-economic-impact-us-horse-industry.

ANAND, Paul, PATTANAIK, Prasanta e PUPPE, Clemens, eds. 2009. *The Oxford Handbook of Rational and Social Choice*. Nova York: Oxford University Press.

ANDRES, B., KOETHE, U., KROEGER, T., HELMSTAEDTER, M., BRIGGMAN, K. L., DENK, W. e HAMPRECHT, F. A. 2012. “3D Segmentation of sbfsem Images of Neuropil by a Graphical Model over Supervoxel Boundaries”. *Medical Image Analysis* 16 (4): 796-805.

ARMSTRONG, Alex. 2012. “Computer Competes in Crossword Tournament”. *I Programmer*, 19 mar.

ARMSTRONG, Stuart. 2007. “Chaining God: A Qualitative Approach to AI, Trust and Moral Systems”. Manuscrito não publicado, 20 out. Recuperado em 31 dez. 2012. Disponível em www.neweuropeancentury.org/GodAI.pdf.

ARMSTRONG, Stuart. 2010. *Utility Indifference*, relatório técnico 2010-1. Oxford: Future of Humanity Institute, University of Oxford.

ARMSTRONG, Stuart. 2013. “General Purpose Intelligence: Arguing the Orthogonality Thesis”. *Analysis and Metaphysics* 12: 68-84.

ARMSTRONG, Stuart e SANDBERG, Anders. 2013. “Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox”. *Acta Astronautica* 89: 1-13.

ARMSTRONG, Stuart e SOTALA, Kaj. 2012. “How We’re Predicting AI — or Failing To”. In *Beyond AI: Artificial Dreams*, editado por Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak e Radek Schuster, 52-75. Pilsen: University of West Bohemia. Recuperado em 2 fev. 2013.

ASIMOV, Isaac. 1942. “Runaround”. *Astounding Science-Fiction*, mar., 94-103.

ASIMOV, Isaac. 1985. *Robots and Empire*. Nova York: Doubleday.

AUMANN, Robert J. 1976. “Agreeing to Disagree”. *Annals of Statistics* 4 (6): 1236-9.

AVERCH, Harvey Allen. 1985. *A Strategic Analysis of Science and Technology Policy*. Baltimore: Johns Hopkins University Press.

AZEVEDO, F. A. C., CARVALHO, L. R. B., GRINBERG, L. T., FARFEL, J. M., FERRETTI, R. E. L., LEITE, R. E. P., JACOB, W., LENT, R. e HERCULANO-HOUZEL, S. 2009. “Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain”. *Journal of Comparative Neurology* 513 (5): 532-41.

BAARS, Bernard J. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Nova York: Oxford University Press.

BARATTA, Joseph Preston. 2004. *The Politics of World Federation: United Nations, UN Reform, Atomic Control*. Westport, CT: Praeger.

BARBER, E. J. W. 1991. *Prehistoric Textiles: The Development of Cloth in the Neolithic and Bronze Ages with Special Reference to the Aegean*. Princeton, NJ: Princeton University Press.

BARTELS, J., ANDREASEN, D., EHIRIM, P., MAO, H., SEIBERT, S., WRIGHT, E. J. e KENNEDY, P. 2008. “Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex”. *Journal of Neuroscience Methods* 174 (2): 168-76.

BARTZ, Jennifer A., ZAKI, Jamil, BOLGER, Niall e OCHSNER, Kevin N. 2011. “Social Effects of Oxytocin in Humans: Context and Person Matter”. *Trends in Cognitive Science* 15 (7): 301-9.

BASTEN, Stuart, LUTZ, Wolfgang e SCHERBOV, Sergei. 2013. “Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility”. *Demographic Research* 28: 1145-66.

BAUM, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge, MA: mit Press.

BAUM, Seth D., GOERTZEL, Ben e GOERTZEL, Ted G. 2011. “How Long Until Human-Level AI? Results from an Expert Assessment”. *Technological Forecasting and Social Change* 78 (1): 185-95.

BEAL, J. e WINSTON, P. 2009. “Guest Editors’ Introduction: The New Frontier of Human-Level Artificial Intelligence”. *IEEE Intelligent Systems* 24 (4): 21-3.

BELL, C. GORDON e GEMMELL, Jim. 2009. *Total Recall: How the E-Memory Revolution Will Change Everything*. Nova York: Dutton.

BENYAMIN, B., POURCAIN, B. St., DAVIS, O. S., DAVIES, G., HANSELL, M. K., BRION, M.-J.A., KIRKPATRICK, R. M., et al. 2013. “Childhood Intelligence is Heritable, Highly Polygenic and Associated With FNBP1L”. *Molecular Psychiatry* (23 jan.).

BERG, Joyce E. e RIETZ, Thomas A. 2003. “Prediction Markets as Decision Support Systems”. *Information Systems Frontiers* 5 (1): 79-93.
BERGER, Theodore W., CHAPIN, J. K., GERHARDT, G. A., SOUSSOU, W. V., TAYLOR, D. M. e TRESCO, P. A., eds. 2008. *Brain-Computer Interfaces: An International Assessment of Research and Development Trends*. Springer.
Berger, T. W., Song, D., Chan, R. H., Marmarelis, V. Z., LaCoss, J., Wills, J., Hampson, R. E., Deadwyler, S. A. e Granacki, J. J. 2012. “A Hippocampal Cognitive Prosthesis: Multi-Input, Multi-Output Nonlinear Modeling and VLSI Implementation”.*IEEE Transactions on Neural Systems and Rehabilitation Engineering* 20 (2): 198-211.
BERLINER, Hans J. 1980a. “Backgammon Computer-Program Beats World Champion”. *Artificial Intelligence* 14 (2): 205-220.
BERLINER, Hans J. 1980b. “Backgammon Program Beats World Champ”. *sigart Newsletter* 69: 6-9.
BERNARDO, José M. e SMITH, Adrian F. M. 1994. *Bayesian Theory*, 1 ed. Wiley Series in Probability & Statistics. Nova York: Wiley.
BIRBAUMER, N., MURGUIALDAY, A. R. e COHEN, L. 2008. “Brain-Computer Interface in Paralysis”. *Current Opinion in Neurology* 21 (6): 634-8.
BIRD, Jon e LAYZELL, Paul. 2002. “The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors”. In *Proceedings of the 2002 Congress on Evolutionary Computation*, 2: 1836-41.
BLAIR, Clay, Jr. 1957. “Passing of a Great Mind: John von Neumann, a Brilliant, Jovial Mathematician, was a Prodigious Servant of Science and His Country”. *Life*, 25 fev., 89-104.
BOBROW, Daniel G. 1968.“Natural Language Input for a Computer Problem Solving System”. In *Semantic Information Processing*, editado por Marvin Minsky, 146-227. Cambridge, MA: mit Press.
BOSTROM, Nick. 1997. “Predictions from Philosophy? How Philosophers Could Make Themselves Useful”. Manuscrito não publicado. Última revisão em 19 set. 1998.
BOSTROM, Nick. 2002a. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Nova York: Routledge.
BOSTROM, Nick. 2002b. “Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards”. *Journal of Evolution and Technology* 9.
BOSTROM, Nick. 2003a. “Are We Living in a Computer Simulation?”. *Philosophical Quarterly* 53 (211): 243-55.
BOSTROM, Nick. 2003b. “Astronomical Waste: The Opportunity Cost of Delayed Technological Development”. *Utilitas* 15 (3): 308-314.
BOSTROM, Nick. 2003c. “Ethical Issues in Advanced Artificial Intelligence”. In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, editado por Iva Smit e George E. Lasker, 2: 12-17. Windsor, ON: International Institute for Advanced Studies in Systems Research/Cybernetics.
BOSTROM, Nick. 2004. “The Future of Human Evolution”. In *Two Hundred Years After Kant, Fifty Years After Turing,* editado por Charles Tandy, 2: 339-371. Death and Anti-Death. Palo Alto, CA: Ria University Press.
BOSTROM, Nick. 2006a. “How Long Before Superintelligence?”. *Linguistic and Philosophical Investigations* 5 (1): 11-30.
BOSTROM, Nick. 2006b. “Quantity of Experience: Brain-Duplication and Degrees of Consciousness”. *Minds and Machines* 16 (2): 185-200.
BOSTROM, Nick. 2006c. “What is a Singleton?”. *Linguistic and Philosophical Investigations* 5 (2): 48-54.
BOSTROM, Nick. 2007. “Technological Revolutions: Ethics and Policy in the Dark”.In *Nanoscale: Issues and Perspectives for the Nano Century*, editado por Nigel M. de S. Cameron e M. Ellen Mitchell, 129-52. Hoboken, NJ: Wiley.
BOSTROM, Nick. 2008a. “Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing”. *MIT Technology Review*, edição de maio/jun., 72-7.
BOSTROM, Nick. 2008b. “Why I Want to Be a Posthuman When I Grow Up”. In *Medical Enhancement and Posthumanity*, editado por Bert Gordijn e Ruth Chadwick, 107-37. Nova York: Springer.
BOSTROM, Nick. 2008c. “Letter from Utopia”. *Studies in Ethics, Law e Technology* 2 (1): 1-7.Bostrom, Nick. 2009a. “Moral Uncertainty — Towards a Solution?”. *Overcoming Bias* (blog), 1 jan.
BOSTROM, Nick. 2009b. “Pascal’s Mugging”. *Analysis* 69 (3): 443-5.
BOSTROM, Nick. 2009c. “The Future of Humanity”. In *New Waves in Philosophy of Technology*, editado por Jan Kyrre Berg Olsen, Evan Selinger e Søren Riis, 186-215. Nova York: Palgrave Macmillan.
BOSTROM, Nick. 2011a. “Information Hazards: A Typology of Potential Harms from Knowledge”. *Review of Contemporary Philosophy* 10: 44-79.
BOSTROM, Nick. 2011b. “Infinite Ethics”. *Analysis and Metaphysics* 10: 9-59.
BOSTROM, Nick. 2012. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents”. In “Theory and Philosophy of AI”, editado por Vincent C. Müller, edição especial, *Minds and Machines* 22 (2): 71-85.
BOSTROM, Nick e ĆIRKOVIĆ, Milan M. 2003. “The Doomsday Argument and the Self-Indication Assumption: Reply to Olum”. *Philosophical Quarterly* 53 (210): 83-91.

BOSTROM, Nick e ORD, Toby. 2006. “The Reversal Test: Eliminating the Status Quo Bias in Applied Ethics”. *Ethics* 116 (4): 656-79.
BOSTROM, Nick e ROACHE, Rebecca. 2011. “Smart Policy: Cognitive Enhancement and the Public Interest”. In *Enhancing Human Capacities*, editado por Julian Savulescu, Ruud ter Meulen e Guy Kahane, 138-49. Malden, MA: Wiley-Blackwell.
BOSTROM, Nick e SANDBERG, Anders. 2009a. “Cognitive Enhancement: Methods, Ethics, Regulatory Challenges”. *Science and Engineering Ethics* 15 (3): 311-41.
BOSTROM, Nick e SANDBERG, Anders. 2009b. “The Wisdom of Nature: An Evolutionary Heuristic for Human Enhancement”. In *Human Enhancement*, 1 ed., editado por Julian Savulescu e Nick Bostrom, 375-416. Nova York: Oxford University Press.
BOSTROM, Nick, SANDBERG, Anders e DOUGLAS, Tom. No prelo. “The Unilateralist’s Curse: The Case for a Principle of Conformity”. *Social Epistemology*.
Bostrom, Nick e Yudkowsky, Eliezer. 2015. “The Ethics of Artificial Intelligence”. In *Cambridge Handbook of Artificial Intelligence*, editado por Keith Frankish e William M. Ramsey, 315–334. Nova York: Cambridge University Press.
BOSWELL, James. 1917. *Boswell’s Life of Johnson.* Nova York: Oxford University Press.
BOUCHARD, T. J. 2004. “Genetic Influence on Human Psychological Traits: A Survey”. *Current Directions in Psychological Science* 13 (4): 148-51.
BOURGET, David e CHALMERS, David. 2009. “The PhilPapers Surveys”. Nov. Disponível em http://philpapers.org/surveys/.
BRADBURY, Robert J. 1999. “Matrioshka Brains”. Versão arquivada. Conforme revisado em 16 ago. 2004. Disponível em http://web.archive.org/web/20090901022552/http://www.aeiveos.com:80/~bradbury/MatrioshkaBrains/.
BRINTON, Crane. 1965. *The Anatomy of Revolution*. Edição revisada. Nova York: Vintage Books.
BRYSON, Arthur E., Jr. e Ho, Yu-Chi. 1969. *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham, MA: Blaisdell.
BUEHLER, Martin, Iagnemma, Karl e Singh, Sanjiv, eds. 2009. *The darpa Urban Challenge: Autonomous Vehicles in City Traffic.* Springer Tracts in Advanced Robotics 56. Berlim: Springer.
BURCH-BROWN, J. 2014. “Clues for Consequentialists”. *Utilitas* 26 (1): 105-19.
BURKE, Colin. 2001. “Agnes Meyer Driscoll vs. the Enigma and the Bombe”. Manuscrito não publicado. Recuperado em 22 fev. 2013. Disponível em http://userpages.umbc.edu/~burke/driscoll1-2011.pdf.
CANBÄCK, S., SAMOUEL, P. e PRICE, D. 2006. “Do Diseconomies of Scale Impact Firm Size and Performance? A Theoretical and Empirical Overview”. *Journal of Managerial Economics* 4 (1): 27-70.
CARMENA, J. M., LEBEDEV, M. A., CRIST, R. E., O’DOHERTY, J. E., SANTUCCI, D. M., DIMITROV, D. F., PATIL, P. G., HENRIQUEZ, C. S. e NICOLELIS, M. A. 2003. “Learning to Control a Brain-Machine Interface for Reaching and Grasping by Primates”. *Public Library of Science Biology* 1 (2): 193-208.
CARROLL, Bradley W. e OSTLIE, Dale A. 2007. An Introduction to Modern Astrophysics. 2 ed. São Francisco: Pearson Addison Wesley.
CARROLL, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Nova York: Cambridge University Press.
CARTER, Brandon. 1983. “The Anthropic Principle and its Implications for Biological Evolution”. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 310 (1512): 347-63.
CARTER, Brandon. 1993. “The Anthropic Selection Principle and the Ultra-Darwinian Synthesis”. In *The Anthropic Principle: Proceedings of the Second Venice Conference on Cosmology and Philosophy*, editado por F. Bertola e U. Curi, 33-66. Cambridge: Cambridge University Press.
CFTC & SEC (Commodity Futures Trading Commission and Securities & Exchange Commission). 2010. *Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. Washington, DC.
CHALMERS, David John. 2010. “The Singularity: A Philosophical Analysis”. *Journal of Consciousness Studies* 17 (9-10): 7-65.
CHASON, R. J., CSOKMAY, J., SEGARS, J. H., DECHERNEY, A. H. e ARMANT, D. R. 2011. “Environmental and Epigenetic Effects Upon Preimplantation Embryo Metabolism and Development”. *Trends in Endocrinology and Metabolism* 22 (10): 412-20.
CHEN, S. e RAVALLION, M. 2010. “The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight Against Poverty”. *Quarterly Journal of Economics* 125 (4): 1577-1625.
CHISLENKO, Alexander. 1996. “Networking in the Mind Age: Some Thoughts on Evolution of Robotics and Distributed Systems”. Manuscrito não publicado.
CHISLENKO, Alexander. 1997. “Technology as Extension of Human Functional Architecture”. *Extropy Online*.
CHOROST, Michael. 2005. *Rebuilt: How Becoming Part Computer Made Me More Human*. Boston: Houghton Mifflin.

CHRISTIANO, Paul F. 2012. "'Indirect Normativity' Write-up". *Ordinary Ideas* (blog), 21 abr.
CIA. 2013. *The World Factbook*. Central Intelligence Agency. Recuperado em 3 ago. Disponível em www.cia.gov/library/publications/the-world-factbook/rankorder/2127rank.html?countryname=United%20States&countrycode=us&regionCode=noa&rank=121#us.
CICERO. 1923. "On Divination". In *On Old Age, on Friendship, on Divination*, trad. W. A. Falconer. Loeb Classical Library. Cambridge, MA: Harvard University Press.
CIRASELLA, Jill e KOPEC, Danny. 2006. "The History of Computer Games". Exibido em Dartmouth Artificial Intelligence Conference: The Next Fifty Years (AI@50), Dartmouth College, 13-15 jul.
ĆIRKOVIĆ, Milan M. 2004. "Forecast for the Next Eon: Applied Cosmology and the Long-Term Fate of Intelligent Beings". *Foundations of Physics* 34 (2): 239-61.
ĆIRKOVIĆ, Milan M., Sandberg, Anders e Bostrom, Nick. 2010. "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks". *Risk Analysis* 30 (10): 1495-1506.
CLARK, Andy e CHALMERS, David J. 1998. "The Extended Mind". *Analysis* 58 (1): 7-19.
CLARK, Gregory. 2007. *A Farewell to Alms: A Brief Economic History of the World.* 1 ed. Princeton, NJ: Princeton University Press.
CLAVIN, Whitney. 2012. "Study Shows Our Galaxy Has at Least 100 Billion Planets". *Jet Propulsion Laboratory*, 11 jan.
CME GROUP. 2010. *What Happened on May 6th?* Chicago, 10 maio.
COASE, R. H. 1937. "The Nature of the Firm". *Economica* 4 (16): 386-405.
COCHRAN, Gregory e HARPENDING, Henry. 2009. *The 10,000 Year Explosion: How Civilization Accelerated Human Evolution*. Nova York: Basic Books.
COCHRAN, G., HARDY, J. e HARPENDING, H. 2006. "Natural History of Ashkenazi Intelligence". *Journal of Biosocial Science* 38 (5): 659-93.
COOK, James Gordon. 1984. *Handbook of Textile Fibres: Natural Fibres*. Cambridge: Woodhead.
COPE, David. 1996. *Experiments in Musical Intelligence*. Computer Music and Digital Audio Series. Madison, WI: A-R Editions.
COTMAN, Carl W. e BERCHTOLD, Nicole C. 2002. "Exercise: A Behavioral Intervention to Enhance Brain Health and Plasticity". *Trends in Neurosciences* 25 (6): 295-301.
COWAN, Nelson. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity". *Behavioral and Brain Sciences* 24 (1): 87-114.
CRABTREE, Steve. 1999. "New Poll Gauges Americans' General Knowledge Levels". *Gallup News*, 6 jul.
CROSS, Stephen E. e WALKER, Edward. 1994. "Dart: Applying Knowledge Based Planning and Scheduling to Crisis Action Planning". In *Intelligent Scheduling*, editado por Monte Zweben e Mark Fox, 711-29. São Francisco, CA: Morgan Kaufmann.
CROW, James F. 2000. "The Origins, Patterns and Implications of Human Spontaneous Mutation". *Nature Reviews Genetics* 1 (1): 40-7.
CYRANOSKI, David. 2013. "Stem Cells: Egg Engineers". *Nature* 500 (7463): 392-4.
DAGNELIE, Gislin. 2012. "Retinal Implants: Emergence of a Multidisciplinary Field". *Current Opinion in Neurology* 25 (1): 67-75.
DAI, Wei. 2009. "Towards a New Decision Theory". *Less Wrong* (blog), 13 ago.
DALRYMPLE, David. 2011. "Comment on Kaufman, J. 'Whole Brain Emulation: Looking at Progress on *C. Elegans*'". *Less Wrong* (blog), 29 out.
DAVIES, G., TENESA, A., PAYTON, A., YANG, J., HARRIS, S. E., LIEWALD, D., Ke, X., et al. 2011. "Genome-Wide Association Studies Establish That Human Intelligence Is Highly Heritable and Polygenic". *Molecular Psychiatry* 16 (10): 996-1005.
DAVIS, Oliver S. P., BUTCHER, Lee M., DOCHERTY, SOPHIA J., MEABURN, EMMA L., CURTIS, Charles J. C., SIMPSON, Michael A., SCHALKWYK, Leonard C. e PLOMIN, Robert. 2010. "A Three-Stage Genome-Wide Association Study of General Cognitive Ability: Hunting the Small Effects". *Behavior Genetics* 40 (6): 759-767.
DAWKINS, Richard. 1995. *River Out of Eden: A Darwinian View of Life*. Science Masters Series. Nova York: Basic Books.
DE BLANC, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems*. Berkeley, CA: Machine Intelligence Research Institute, 19 maio.
DE LONG, J. Bradford. 1998. "Estimates of World gdp, One Million B.C.-Present". Manuscrito não publicado.
DE RAEDT, Luc e FLACH, Peter, eds. 2001. *Machine Learning: ecml 2001: 12th European Conference on Machine Learning, Freiburg, Germany, September 5-7, 2001. Proceedings*. Lecture Notes in Computer Science 2167. Nova York: Springer.
DEAN, Cornelia. 2005. "Scientific Savvy? In U.S., Not Much". *New York Times,* 30 ago.
DEARY, Ian J. 2001. "Human Intelligence Differences: A Recent History". *Trends in Cognitive Sciences* 5 (3): 127-30.

DEARY, Ian J. 2012. “Intelligence”. *Annual Review of Psychology* 63: 453-82.
DEARY, Ian J., PENKE, L. e JOHNSON, W. 2010. “The Neuroscience of Human Intelligence Differences”. *Nature Reviews Neuroscience* 11 (3): 201-11.
DEGNAN, G. G., WIND, T. C., JONES, E. V. e EDLICH, R. F. 2002. “Functional Electrical Stimulation in Tetraplegic Patients to Restore Hand Function”. *Journal of Long-Term Effects of Medical Implants* 12 (3): 175-88.
DEVLIN, B., DANIELS, M. e ROEDER, K. 1997. “The Heritability of IQ”. *Nature* 388 (6641): 468-71.
DEWEY, Daniel. 2011. “Learning What to Value”. In *Artificial General Intelligence: 4th International Conference, agi 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings,* editado por Jürgen Schmidhuber, Kristinn R. Thórisson e Moshe Looks, 309-14. Lecture Notes in Computer Science 6830. Berlim: Springer.
DOWE, D. L. e HERNÁNDEZ-ORALLO, J. 2012. “IQ Tests Are Not for Machines, Yet”. *Intelligence* 40 (2): 77-81.
DRESCHER, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics*. Bradford Books. Cambridge, MA: mit Press.
DREXLER, K. Eric. 1986. *Engines of Creation*. Garden City, NY: Anchor.
DREXLER, K. Eric. 1992. *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. Nova York: Wiley.
DREXLER, K. Eric. 2013. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*. Nova York: PublicAffairs.
DRISCOLL, Kevin. 2012. “Code Critique: ‘Altair Music of a Sort’”. Artigo apresentado na Critical Code Studies Working Group, conferência on-line, 6 fev.
DYSON, Freeman J. 1960. “Search for Artificial Stellar Sources of Infrared Radiation”. *Science* 131 (3414): 1667-1668.
DYSON, Freeman J. 1979. *Disturbing the Universe*. 1 ed. Sloan Foundation Science Series. Nova York: Harper & Row.
ELGA, Adam. 2004. “Defeating Dr. Evil with Self-Locating Belief”. *Philosophy and Phenomenological Research* 69 (2): 383-96.
ELGA, Adam. 2007. “Reflection and Disagreement”. *Noûs* 41 (3): 478-502.
ELIASMITH, Chris, STEWART, Terrence C., CHOO, Xuan, BEKOLAY, Trevor, DEWOLF, Travis, TANG, Yichuan e RASMUSSEN, Daniel. 2012. “A Large-Scale Model of the Functioning Brain”. *Science* 338 (6111): 1202-5.
ELLIS, J. H. 1999. “The History of Non-Secret Encryption”. *Cryptologia* 23 (3): 267-73.
ELYASAF, Achiya, HAUPTMANN, Ami e SIPPER, Moche. 2011. “Ga-Freecell: Evolving Solvers for the Game of Freecell”. In *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference*, 1931-1938. GECCO’ 11. Nova York: acm.
EPPIG, C., FINCHER, C. L. e THORNHILL, R. 2010. “Parasite Prevalence and the Worldwide Distribution of Cognitive Ability”. *Proceedings of the Royal Society B: Biological Sciences* 277 (1701): 3801-8.
ESPENSHADE, T. J., GUZMAN, J. C. e WESTOFF, C. F. 2003. “The Surprising Global Variation in Replacement Fertility”. *Population Research and Policy Review* 22 (5-6): 575-83.
EVANS, Thomas G. 1964. “A Heuristic Program to Solve Geometric-Analogy Problems”. In *Proceedings of the April 21-23, 1964, Spring Joint Computer Conference*, 327-338. afips ’64. Nova York: acm.
EVANS, Thomas G. 1968. “A Program for the Solution of a Class of Geometric-Analogy Intelligence-Test Questions”. In *Semantic Information Processing*, editado por Marvin Minsky, 271-353. Cambridge, MA: mit Press.
FAISAL, A. A., SELEN, L. P. e WOLPERT, D. M. 2008. “Noise in the Nervous System”. *Nature Reviews Neuroscience* 9 (4): 292-303.
FAISAL, A. A., WHITE, J. A. e LAUGHLIN, S. B. 2005. “Ion-Channel Noise Places Limits on the Miniaturization of the Brain’s Wiring”. *Current Biology* 15 (12): 1143-9.
FELDMAN, Jacob. 2000. “Minimization of Boolean Complexity in Human Concept Learning”. *Nature* 407 (6804): 630-3.
FELDMAN, J. A. e BALLARD, Dana H. 1982. “Connectionist Models and Their Properties”. *Cognitive Science* 6 (3): 205-254.
FOLEY, J. A., MONFREDA, C., RAMANKUTTY, N. e ZAKS, D. 2007. “Our Share of the Planetary Pie”. *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12585-6.
FORGAS, Joseph P., COOPER, Joel E CRANO, William D., eds. 2010. *The Psychology of Attitudes and Attitude Change*. Sydney Symposium of Social Psychology. Nova York: Psychology Press.
FRANK, Robert H. 1999. *Luxury Fever: Why Money Fails to Satisfy in an Era of Excess*. Nova York: Free Press.
FREDRIKSEN, Kaja Bonesmo. 2012. *Less Income Inequality and More Growth — Are They Compatible?: Part 6. The Distribution of Wealth*. Relatório técnico, oecd Economics Department Working Papers 929. oecd Publishing.
FREITAS, Robert A., Jr. 1980. “A Self-Replicating Interstellar Probe”. *Journal of the British Interplanetary Society* 33: 251-64.

FREITAS, Robert A., Jr. 2000. “Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations”. Foresight Institute. Abril. Recuperado em 28 jul. 2013. Disponível em www.foresight.org/nano/Ecophagy.html.
FREITAS, Robert A., Jr. e MERKLE, Ralph C. 2004. *Kinematic Self-Replicating Machines*. Georgetown, TX: Landes Bioscience.
GADDIS, John Lewis. 1982. *Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy*. Nova York: Oxford University Press.
GAMMONED.NET. 2012. “Snowie”. Versão arquivada. Recuperado em jun. 30. Disponível em http://web.archive.org/web/20070920191840/http://www.gammoned.com/snowie.html.
GATES, Bill. 1975. “Software Contest Winners Announced”. *Computer Notes* 1 (2): 1.
Georgieff, Michael K. 2007. “Nutrition and the Developing Brain: Nutrient Priorities and Measurement”. *American Journal of Clinical Nutrition* 85 (2): 614S-620S.
GIANAROLI, Luca. 2000. “Preimplantation Genetic Diagnosis: Polar Body and Embryo Biopsy”. Suplemento, *Human Reproduction* 15 (4): 69-75.
GILOVICH, Thomas, Griffin, Dale e Kahneman, Daniel, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Nova York: Cambridge University Press.
GILSTER, Paul. 2012. “ESO: Habitable Red Dwarf Planets Abundant”. *Centauri Dreams* (blog), 29 mar.
GOLDSTONE, Jack A. 1980. “Theories of Revolution: The Third Generation”. *World Politics* 32 (3): 425-53.
GOLDSTONE, Jack A. 2001. “Towards a Fourth Generation of Revolutionary Theory”. *Annual Review of Political Science* 4: 139-87.
GOOD, Irving John. 1965. “Speculations Concerning the First Ultraintelligent Machine”. In *Advances in Computers*, editado por Franz L. Alt e Morris Rubinoff, 6: 31-88. Nova York: Academic Press.
GOOD, Irving John. 1970. “Some Future Social Repercussions of Computers”. *International Journal of Environmental Studies* 1 (1-4): 67-79.
GOOD, Irving John. 1976. “Book review of ‘The Thinking Computer: Mind Inside Matter’”. In *International Journal of Man-Machine Studies* 8: 617-20.
GOOD, Irving John. 1982. “Ethical Machines”. In *Intelligent Systems: Practice and Perspective*, editado por J. E. Hayes, Donald Michie e Y.-H. Pao, 555-60. Machine Intelligence 10. Chichester: Ellis Horwood.
GOODMAN, Nelson. 1954. *Fact, Fiction, and Forecast*. 1 ed. Londres: Athlone Press.
GOTT, J. R., Juric, M., Schlegel, D., Hoyle, F., Vogeley, M., Tegmark, M., Bahcall, N. e Brinkmann, J. 2005. “A Map of the Universe”. *Astrophysical Journal* 624 (2): 463-83.
GOTTFREDSON, Linda S. 2002. “G: Highly General and Highly Practical”. In *The General Factor of Intelligence: How General Is It?*, editado por Robert J. Sternberg e Elena L. Grigorenko, 331-80. Mahwah, NJ: Lawrence Erlbaum.
GOULD, S. J. 1990. *Wonderful Life: The Burgess Shale and the Nature of History*. Nova York: Norton.
GRAHAM, Gordon. 1997. *The Shape of the Past: A Philosophical Approach to History*. Nova York: Oxford University Press.
GRAY, C. M. e MCCORMICK, D. A. 1996. “Chattering Cells: Superficial Pyramidal Neurons Contributing to the Generation of Synchronous Oscillations in the Visual Cortex”. *Science* 274 (5284): 109-13.
GREENE, Kate. 2012. “Intel’s Tiny Wi-Fi Chip Could Have a Big Impact”. *mit Technology Review*, 21 set.
GUIZZO, Erico. 2010. “World Robot Population Reaches 8.6 Million”. *ieee Spectrum*, 14 abr.
GUNN, James E. 1982. *Isaac Asimov: The Foundations of Science Fiction*. Science-Fiction Writers. Nova York: Oxford University Press.
HABERL, Helmut, ERB, Karl-Heinz e KRAUSMANN, Fridolin. 2013. “Global Human Appropriation of Net Primary Production (hanpp)”. *Encyclopedia of Earth*, 3 set.
HABERL, H., ERB, K. H., KRAUSMANN, F., GAUBE, V., BONDEAU, A., PLUTZAR, C., GINGRICH, S., LUCHT, W. e FISCHER-KOWALSKI, M. 2007. “Quantifying and Mapping the Human Appropriation of Net Primary Production in Earth’s Terrestrial Ecosystems”. *Proceedings of the National Academy of Sciences of the United States of America* 104 (31): 12942-7.
HÁJEK, Alan. 2009. “Dutch Book Arguments”. In Anand, Paul, Pattanaik e Puppe 2009 173-95.
HALL, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine*. Amherst, NY: Prometheus Books.
HAMPSON, R. E., SONG, D., CHAN, R. H., SWEATT, A. J., RILEY, M. R., GERHARDT, G. A., SHIN, D. C., MARMARELIS, V. Z., BERGER, T. W. e DEADWYLER, S. A. 2012. “A Nonlinear Model for Hippocampal Cognitive Prosthesis: Memory Facilitation by Hippocampal Ensemble Stimulation”. *ieee Transactions on Neural Systems and Rehabilitation Engineering* 20 (2): 184-97.
HANSON, Robin. 1994. “If Uploads Come First: The Crack of a Future Dawn”. *Extropy* 6 (2).
HANSON, Robin. 1995. “Could Gambling Save Science? Encouraging an Honest Consensus”. *Social Epistemology* 9 (1): 3-33.

HANSON, Robin. 1998a. “Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization”. Manuscrito não publicado, 1 jul. Recuperado em 26 abr. 2012. Disponível em http://hanson.gmu.edu/filluniv.pdf.
HANSON, Robin. 1998b. “Economic Growth Given Machine Intelligence”. Manuscrito não publicado. Recuperado em 15 maio 2013. Disponível em http://hanson.gmu.edu/aigrow.pdf.
HANSON, Robin. 1998c. “Long-Term Growth as a Sequence of Exponential Modes”. Manuscrito não publicado. Última revisão em dez. 2000. Disponível em http://hanson.gmu.edu/longgrow.pdf.
HANSON, Robin. 1998d. “Must Early Life Be Easy? The Rhythm of Major Evolutionary Transitions”. Manuscrito não publicado, 23 set. Recuperado em 12 ago. 2012. Disponível em http://hanson.gmu.edu/hardstep.pdf.
HANSON, Robin. 2000. “Shall We Vote on Values, But Bet on Beliefs?”. Manuscrito não publicado, set. Última revisão em out. 2007. Disponível em http://hanson.gmu.edu/futarchy.pdf.
HANSON, Robin. 2006. “Uncommon Priors Require Origin Disputes”. *Theory and Decision* 61 (4): 319-328.
HANSON, Robin. 2008. “Economics of the Singularity”. *IEEE Spectrum* 45 (6): 45-50.
HANSON, Robin. 2009. “Tiptoe or Dash to Future?”. *Overcoming Bias* (blog), 23 dez.
HANSON, Robin. 2012. “Envisioning the Economy, and Society, of Whole Brain Emulations”. Artigo apresentado na conferência agi Impacts, Oxford, 8-11 dez.
HART, Oliver. 2008. “Economica Coase Lecture Reference Points and the Theory of the Firm”. *Economica* 75 (299): 404-11.
HAY, Nicholas James. 2005. “Optimal Agents”. Tese de bacharelado, University of Auckland.
HEDBERG, Sara Reese. 2002. “Dart: Revolutionizing Logistics Planning”. *IEEE Intelligent Systems* 17 (3): 81-3.
HELLIWELL, John, LAYARD, Richard e SACHS, Jeffrey. 2012. *World Happiness Report*. The Earth Institute.
HELMSTAEDTER, M., BRIGGMAN, K. L. e DENK, W. 2011. “High-Accuracy Neurite Reconstruction for High-Throughput Neuroanatomy”. *Nature Neuroscience* 14 (8): 1081-8.
HEYL, Jeremy S. 2005. “The Long-Term Future of Space Travel”. *Physical Review D* 72 (10): 1-4.
HIBBARD, Bill. 2011. “Measuring Agent Intelligence via Hierarchies of Environments”. In *Artificial General Intelligence: 4th International Conference, agi 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, editado por Jürgen Schmidhuber, Kristinn R. Thórisson e Moshe Looks, 303-8. Lecture Notes in Computer Science 6830. Berlim: Springer.
HINKE, R. M., HU, X., STILLMAN, A. E., HERKLE, H., SALMI, R. e UGURBIL, K. 1993. “Functional Magnetic Resonance Imaging of Broca’s Area During Internal Speech”. *Neuroreport* 4 (6): 675-8.
HINXTON GROUP. 2008. *Consensus Statement: Science, Ethics and Policy Challenges of Pluripotent Stem Cell-Derived Gametes*. Hinxton, Cambridgeshire, UK, 11 abr. Disponível em www.hinxtongroup.org/Consensus_HG08_FINAL.pdf.
HOFFMAN, David E. 2009. *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy*. Nova York: Doubleday.
HOFSTADTER, Douglas R. (1979) 1999. *Gödel, Escher, Bach: An Eternal Golden Braid*. Nova York: Basic Books.
HOLLEY, Rose. 2009. “How Good Can It Get? Analysing and Improving ocr Accuracy in Large Scale Historic Newspaper Digitisation Programs”. *D-Lib Magazine* 15 (3-4).
HORTON, Sue, Alderman, Harold e Rivera, Juan A. 2008. *Copenhagen Consensus 2008 Challenge Paper: Hunger and Malnutrition*. Relatório técnico. Copenhagen Consensus Center, 11 maio.
HOWSON, Colin e URBACH, Peter. 1993. *Scientific Reasoning: The Bayesian Approach*. 2 ed. Chicago: Open Court.
HSU, Stephen. 2012. “Investigating the Genetic Basis for Intelligence and Other Quantitative Traits”. Palestra ministrada no UC Davis Department of Physics Colloquium, Davis, CA, 13 fev.
HUEBNER, Bryce. 2008. “Do You See What We See? An Investigation of an Argument Against Collective Representation”. *Philosophical Psychology* 21 (1): 91-112.
HUFF, C. D., XING, J., ROGERS, A. R., WITHERSPOON, D. e JORDE, L. B. 2010. “Mobile Elements Reveal Small Population Size in the Ancient Ancestors of *Homo Sapiens*”. *Proceedings of the National Academy of Sciences of the United States of America* 107 (5): 2147-52.
HUFFMAN, W. Cary e PLESS, Vera. 2003. *Fundamentals of Error-Correcting Codes*. Nova York: Cambridge University Press.
HUNT, Patrick. 2011. “Late Roman Silk: Smuggling and Espionage in the 6th Century CE”. *Philolog, Stanford University* (blog), 2 ago.
HUTTER, Marcus. 2001. “Towards a Universal Theory of Artificial Intelligence Based on Algorithmic Probability and Sequential Decisions”. In De Raedt e Flach 2001, 226-38.
Hutter, Marcus. 2005. *Universal Artificial Intelligencet: Sequential Decisions Based On Algorithmic Probability*. Texts in Theoretical Computer Science. Berlim: Springer.
ILIADOU, A. N., JANSON, P. C. e CNATTINGIUS, S. 2011. “Epigenetics and Assisted Reproductive Technology”. *Journal of Internal Medicine* 270 (5): 414-20.

ISAKSSON, Anders. 2007. *Productivity and Aggregate Growth: A Global Picture*. Relatório técnico 05/2007. Viena, Áustria: unido (United Nations Industrial Development Organization) Research and Statistics Branch.
JONES, Garret. 2009. “Artificial Intelligence and Economic Growth: A Few Finger-Exercises”. Manuscrito não publicado, jan. Recuperado em 5 nov. 2012. Disponível em http://mason.gmu.edu/~gjonesb/AIandGrowth.
Jones, Vincent C. 1985. *Manhattan: The Army and the Atomic Bomb*. United States Army in World War ii. Washington, DC: Center of Military History.
Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction and Decision Theory. Nova York: Cambridge University Press.
JUDD, K. L., SCHMEDDERS, K. e YELTEKIN, S. 2012. “Optimal Rules for Patent Races”. *International Economic Review* 53 (1): 23-52.
KALFOGLOU, A., SUTHERS, K., SCOTT, J. e HUDSON, K. 2004. *Reproductive Genetic Testing: What America Thinks*. Genetics and Public Policy Center.
KAMM, Frances M. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford Ethics Series. Nova York: Oxford University Press.
KANDEL, Eric R., SCHWARTZ, James H. e JESSELL, Thomas M., eds. 2000. *Principles of Neural Science*. 4 ed. Nova York: McGraw-Hill.
KANSA, Eric. 2003. “Social Complexity and Flamboyant Display in Competition: More Thoughts on the Fermi Paradox”. Manuscrito não publicado, versão arquivada.
KARNOFSKY, Holden. 2012. “Comment on ‘Reply to Holden on Tool AI’”. *Less Wrong* (blog), 1 ago.
KASPAROV, Garry. 1996. “The Day That I Sensed a New Kind of Intelligence”. *Time*, 25 mar., n. 13.
KAUFMAN, Jeff. 2011. “Whole Brain Emulation and Nematodes”. *Jeff Kaufman’s Blog* (blog), 2 nov.
KEIM, G. A., SHAZEER, N. M., LITTMAN, M. L., AGARWAL, S., CHEVES, C. M., FITZGERALD, J., GROSLAND, J., JIANG, F., POLLARD, S. e WEINMEISTER, K. 1999. “Proverb: The Probabilistic Cruciverbalist”. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 710-17. Menlo Park, CA: aaai Press.
KELL, Harrison J., LUBINSKI, David e BENBOW, Camilla P. 2013. “Who Rises to the Top? Early Indicators”. *Psychological Science* 24 (5): 648-59.
KELLER, Wolfgang. 2004. “International Technology Diffusion”. *Journal of Economic Literature* 42 (3): 752-82.
KGS GO SERVER. 2012. “KGS Game Archives: Games of KGS player zen19”. Recuperado em 22 jul. 2013. Disponível em www.gokgs.com/gameArchives.jsp?user=zen19d&oldAccounts=t&year=2012&month=3.
KNILL, Emanuel, LAFLAMME, Raymond e VIOLA, Lorenza. 2000. “Theory of Quantum Error Correction for General Noise”. *Physical Review Letters* 84 (11): 2525-8.
KOCH, K., MCLEAN, J., SEGEV, R., FREED, M. A., BERRY, M. J., BALASUBRAMANIAN, V. e STERLING, P. 2006. “How *Much* the Eye Tells the Brain”. *Current Biology* 16 (14): 1428-34.
KONG, A., FRIGGE, M. L., MASSON, G., BESENBACHER, S., SULEM, P., MAGNUSSON, G., GUDJONSSON, S. A., SIGURDSSON, A., et al. 2012. “Rate of De Novo Mutations and the Importance of Father’s Age to Disease Risk”. *Nature* 488: 471-5.
KOOMEY, Jonathan G. 2011. *Growth in Data Center Electricity Use 2005 to 2010*. Relatório técnico, 8/01/2011. Oakland, CA: Analytics Press.
KOUBI, Vally. 1999. “Military Technology Races”. *International Organization* 53 (3): 537-65.
KOUBI, Vally e LALMAN, David. 2007. “Distribution of Power and Military R&D”. *Journal of Theoretical Politics* 19 (2): 133-52.
KOZA, J. R., KEANE, M. A., STREETER, M. J., MYDLOWEC, W., YU, J. e LANZA, G. 2003. *Genetic Programming iv: Routine Human-Competitive Machine Intelligence*. 2 ed. Genetic Programming. Norwell, MA: Kluwer Academic.
KREMER, Michael. 1993. “Population Growth and Technological Change: One Million B.C. to 1990”. *Quarterly Journal of Economics* 108 (3): 681-716.
KRUEL, Alexander. 2011. “Interview Series on Risks from AI”. *Less Wrong Wiki* (blog). Recuperado em 26 out. 2013. Disponível em http://wiki.lesswrong.com/wiki/Interview_series_on_risks_from_AI.
KRUEL, Alexander. 2012. “Q&A with Experts on Risks From AI #2”. *Less Wrong* (blog), 9 jan.
KRUSIENSKI, D. J. e SHIH, J. J. 2011. “Control of a Visual Keyboard Using an Electrocorticographic Brain-Computer Interface”. *Neurorehabilitation and Neural Repair* 25 (4): 323-31.
KUHN, Thomas S. 1962. *The Structure of Scientific Revolutions*. 1 ed. Chicago: University of Chicago Press.
Kuipers, Benjamin. 2012. “An Existing, Ecologically-Successful Genus of Collectively Intelligent Artificial Creatures”. Artigo apresentado na 4th International Conference, iccci 2012, Ho Chi Minh City, Vietnã, 28-30 nov.
KURZWEIL, Ray. 2001. “Response to Stephen Hawking”. Kurzweil Accelerating Intelligence. 5 set. Recuperado em 31 dez. 2012. Disponível em www.kurzweilai.net/response-to-stephen-hawking.
KURZWEIL, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. Nova York: Viking.
LAFFONT, Jean-Jacques e MARTIMORT, David. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton, NJ: Princeton University Press.
Lancet, The. 2008. “Iodine Deficiency — Way to Go Yet”. *The Lancet* 372 (9633): 88.

LANDAUER, Thomas K. 1986. “How Much Do People Remember? Some Estimates of the Quantity of Learned Information in Long-Term Memory”. *Cognitive Science* 10 (4): 477-93.
LEBEDEV, Anastasiya. 2004. “The Man Who Saved the World Finally Recognized”. *MosNews*, 21 maio.
LEBEDEV, M. A. e NICOLELIS, M. A. 2006. “Brain-Machine Interfaces: Past, Present and Future”. *Trends in Neuroscience* 29 (9): 536-46.
LEGG, Shane. 2008. “Machine Super Intelligence”. Dissertação de doutorado, University of Lugano.
LEIGH, E. G., Jr. 2010. “The Group Selection Controversy”. *Journal of Evolutionary Biology* 23(1): 6-19.
LENAT, Douglas B. 1982. “Learning Program Helps Win National Fleet Wargame Tournament”. *sigart Newsletter* 79: 16-17.
Lenat, Douglas B. 1983. “EURISKO: A Program that Learns New Heuristics and Domain Concepts”. *Artificial Intelligence* 21 (1-2): 61-98.
LENMAN, James. 2000. “Consequentialism and Cluelessness”. *Philosophy & Public Affairs* 29 (4): 342-70.
LERNER, Josh. 1997. “An Empirical Exploration of a Technology Race”. *RAND Journal of Economics* 28 (2): 228-47.
Leslie, John. 1996. *The End of the World: The Science and Ethics of Human Extinction*. Londres: Routledge.
LEWIS, David. 1988. “Desire as Belief”. *Mind: A Quarterly Review of Philosophy* 97 (387): 323-32.
LI, Ming e VITAÁNYI, Paul M. B. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Nova York: Springer.
LIN, Thomas, MAUSAM e ETZIONI, Oren. 2012. “Entity Linking at Web Scale”. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX ’12)*, editado por James Fan, Raphael Hoffman, Aditya Kalyanpur, Sebastian Riedel, Fabian Suchanek e Partha Pratim Talukdar, 84-88. Madison, WI: Omnipress.
LLOYD, Seth. 2000. “Ultimate Physical Limits to Computation”. *Nature* 406 (6799): 1047-54.
LOUIS HARRIS & ASSOCIATES. 1969. “Science, Sex, and Morality Survey, study n. 1927”. *Life Magazine* (Nova York) 4.
LYNCH, Michael. 2010. “Rate, Molecular Spectrum, and Consequences of Human Mutation”. *Proceedings of the National Academy of Sciences of the United States of America* 107 (3): 961-8.
LYONS, Mark K. 2011. “Deep Brain Stimulation: Current and Future Clinical Applications”. *Mayo Clinic Proceedings* 86 (7): 662-72.
MACASKILL, William. 2010. “Moral Uncertainty and Intertheoretic Comparisons of Value”. Tese de bacharelado em filosofia, University of Oxford.
MCCARTHY, John. 2007. “From Here to Human-Level AI”. *Artificial Intelligence* 171 (18): 1174-82.
MCCORDUCK, Pamela. 1979. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. São Francisco: W. H. Freeman.
MACK, C. A. 2011. “Fifty Years of Moore’s Law”. *ieee Transactions on Semiconductor Manufacturing* 24 (2): 202-7.
MACKAY, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms*. Nova York: Cambridge University Press.
MCLEAN, George e STEWART, Brian. 1979. “Norad False Alarm Causes Uproar”. *The National*. Transmitido em 10 nov. Ottawa, ON: cbc, 2012. Transmissão de notícia.
MADDISON, Angus. 1999. “Economic Progress: The Last Half Century in Historical Perspective”. In *Facts and Fancies of Human Development: Annual Symposium and Cunningham Lecture, 1999*, editado por Ian Castles. Occasional Paper Series, 1/2000. Acton, ACT: Academy of the Social Sciences in Australia.
MADDISON, Angus. 2001. *The World Economy: A Millennial Perspective*. Development Centre Studies. Paris: Development Centre of the Organisation for Economic Co-operation/Development.
MADDISON, Angus. 2005. *Growth and Interaction in the World Economy: The Roots of Modernity*. Washington, DC: aei Press.
MADDISON, Angus. 2007. *Contours of the World Economy, 1-2030 AD: Essays in Macro-Economic History*. Nova York: Oxford University Press.
MADDISON, Angus. 2010. “Statistics of World Population, gdp and Per Capita gdp 1-2008 AD”. Recuperado em 26 out. 2013. Disponível em www.ggdc.net/maddison/Historical_Statistics/vertical-file_02-2010.xls.
MAI, Q., YU, Y., LI, T., WANG, L., CHEN, M. J., HUANG, S. Z., ZHOU, C. e ZHOU, Q. 2007. “Derivation of Human Embryonic Stem Cell Lines from Parthenogenetic Blastocysts”. *Cell Research* 17 (12): 1008-19.
MAK, J. N. e WOLPAW, J. R. 2009. “Clinical Applications of Brain-Computer Interfaces: Current State and Future Prospects”. *ieee Reviews in Biomedical Engineering* 2: 187-99.
MANKIW, N. Gregory. 2009. *Macroeconomics*. 7 ed. Nova York, NY: Worth.
MARDIS, Elaine R. 2011. “A Decade’s Perspective on dna Sequencing Technology”. *Nature* 470 (7333): 198-203.
MARKOFF, John. 2011. “Computer Wins on ‘Jeopardy!’: Trivial, It’s Not”. *New York Times*, 16 fev.
MARKRAM, Henry. 2006. “The Blue Brain Project”. *Nature Reviews Neuroscience* 7 (2): 153-160.

MASON, Heather. 2003. “Gallup Brain: The Birth of In Vitro Fertilization”. *Gallup*, 5 ago.
MENZEL, Randolf e GIURFA, Martin. 2001. “Cognitive Architecture of a Mini-Brain: The Honeybee”. *Trends in Cognitive Sciences* 5 (2): 62-71.
METZINGER, Thomas. 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: mit Press.
MIJIC, Roko. 2010. “Bootstrapping Safe agi Goal Systems”. Artigo apresentado no workshop Roadmaps to agi and the Future of agi, Lugano, Suíça, 8 mar.
MIKE, Mike. 2013. “Face of Tomorrow”. Recuperado em 30 jun. 2012. Disponível em http://faceoftomorrow.org.
MILGROM, Paul e ROBERTS, John. 1990. “Bargaining Costs, Influence Costs, and the Organization of Economic Activity”. In *Perspectives on Positive Political Economy*, editado por James E. Alt e Kenneth A. Shepsle, 57-89. Nova York: Cambridge University Press.
MILLER, George A. 1956. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information”. *Psychological Review* 63 (2): 81-97.
MILLER, Geoffrey. 2000. *The Mating Mind: How Sexual Choice Shaped the Evolution of Human Nature*. Nova York: Doubleday.
MILLER, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World*. Dallas, TX: BenBella Books.
MINSKY, Marvin. 1967. *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice-Hall.
MINSKY, Marvin, ed. 1968. *Semantic Information Processing*. Cambridge, MA: mit Press.
MINSKY, Marvin. 1984. “Afterword to Vernor Vinge’s novel, ‘True Names’”. Manuscrito não publicado, 1 out. Recuperado em 31 dez. 2012. Disponível em http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html.
MINSKY, Marvin. 2006. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Nova York: Simon & Schuster.
MINSKY, Marvin e PAPERT, Seymour. 1969. *Perceptrons: An Introduction to Computational Geometry*. 1 ed. Cambridge, MA: mit Press.
MOORE, Andrew. 2011. “Hedonism”. In *The Stanford Encyclopedia of Philosophy*, inverno 2011, editado por Edward N. Zalta. Stanford, CA: Stanford University.
MORAVEC, Hans P. 1976. “The Role of Raw Power in Intelligence”. Manuscrito não publicado, 12 maio. Recuperado em 12 ago. 2012. Disponível em www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/l975/Raw.Power.html.
MORAVEC, Hans P. 1980. “Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover”. Dissertação de doutorado, Stanford University.
MORAVEC, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence*. Cambridge, MA: Harvard University Press.
MORAVEC, Hans P. 1998. “When Will Computer Hardware Match the Human Brain?”. *Journal of Evolution and Technology* 1.
MORAVEC, Hans P. 1999. “Rise of the Robots”. *Scientific American*, dez., 124-35.
MUEHLHAUSER, Luke e HELM, Louie. 2012. “The Singularity and Machine Ethics”. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, editado por Amnon Eden, Johnny Søraker, James H. Moor e Eric Steinhart. The Frontiers Collection. Berlim: Springer.
MUEHLHAUSER, Luke e SALAMON, Anna. 2012. “Intelligence Explosion: Evidence and Import”. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, editado por Amnon Eden, Johnny Søraker, James H. Moor e Eric Steinhart. The Frontiers Collection. Berlim: Springer.
MÜLLER, Vincent C. e BOSTROM, Nick. 2016. “Future Progress in Artificial Intelligence: A Survey of Expert Opinion”. In *Fundamental Issues of Artificial Intelligence*, editado por Vincent C. Müller. Synthese Library; Berlim: Springer.
MURPHY, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. Cambridge, MA: mit Press.
NACHMAN, Michael W. e CROWELL, Susan L. 2000. “Estimate of the Mutation Rate per Nucleotide in Humans”. *Genetics* 156 (1): 297-304.
NAGY, Z. P. e CHANG, C. C. 2007. “Artificial Gametes”. *Theriogenology* 67 (1): 99-104.
NAGY, Z. P., KERKIS, I. e CHANG, C. C. 2008. “Development of Artificial Gametes”. *Reproductive BioMedicine Online* 16 (4): 539-44.
NASA. 2013. “International Space Station: Facts and Figures”. Disponível em www.nasa.gov/worldbook/intspacestation_worldbook.html.
NEWBORN, Monty. 2011. *Beyond Deep Blue: Chess in the Stratosphere*. Nova York: Springer.
NEWELL, Allen, SHAW, J. C. e SIMON, Herbert A. 1958. “Chess-Playing Programs and the Problem of Complexity”. *IBM Journal of Research and Development* 2 (4): 320-35.

NEWELL, Allen, SHAW, J. C. e SIMON, Herbert A. 1959. “Report on a General Problem-Solving Program: Proceedings of the International Conference on Information Processing”. In *Information Processing*, 256-64. Paris: Unesco.

NICOLELIS, Miguel A. L. e LEBEDEV, Mikhail A. 2009. “Principles of Neural Ensemble Physiology Underlying the Operation of Brain-Machine Interfaces”. *Nature Reviews Neuroscience* 10 (7): 530-40.

NILSSON, Nils J. 1984. *Shakey the Robot*, Technical Note 323. Menlo Park, CA: AI Center, SRI International, abr.

NILSSON, Nils J. 2009. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Nova York: Cambridge University Press.

NISBETT, R. E., ARONSON, J., BLAIR, C., DICKENS, W., FLYNN, J., HALPERN, D. F. e TURKHEIMER, E. 2012. “Intelligence: New Findings and Theoretical Developments”. *American Psychologist* 67 (2): 130-59.

NIVEN, Larry. 1973. “The Defenseless Dead”. In *Ten Tomorrows,* editado por Roger Elwood, 91-142. Nova York: Fawcett.

NORDHAUS, William D. 2007. “Two Centuries of Productivity Growth in Computing”. *Journal of Economic History* 67 (1): 128-59.

Norton, John D. 2011. “Waiting for Landauer”. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42 (3): 184-98.

OLDS, James e MILNER, Peter. 1954. “Positive Reinforcement Produced by Electrical Stimulation of Septal Area and Other Regions of Rat Brain”. *Journal of Comparative and Physiological Psychology* 47 (6): 419-27.

OLUM, Ken D. 2002. “The Doomsday Argument and the Number of Possible Observers”. *Philosophical Quarterly* 52 (207): 164-84.

OMOHUNDRO, Stephen M. 2007. “The Nature of Self-Improving Artificial Intelligence”. Artigo apresentado na conferência Singularity Summit 2007, São Francisco, CA, 8-9 set.

OMOHUNDRO, Stephen M. 2008. “The Basic AI Drives”. In *Artificial General Intelligence 2008: Proceedings of the First agi Conference*, editado por Pei Wang, Ben Goertzel e Stan Franklin, 483-92. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

OMOHUNDRO, Stephen M. 2012. “Rational Artificial Intelligence for the Greater Good”. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*, editado por Amnon Eden, Johnny Søraker, James H. Moor e Eric Steinhart. The Frontiers Collection. Berlim: Springer.

O'NEILL, Gerard K. 1974. “The Colonization of Space”. *Physics Today* 27 (9): 32-40.

OSHIMA, Hideki e KATAYAMA, Yoichi. 2010. “Neuroethics of Deep Brain Stimulation for Mental Disorders: Brain Stimulation Reward in Humans”. *Neurologia medico-chirurgica* 50 (9): 845-52.

PARFIT, Derek. 1986. *Reasons and Persons*. Nova York: Oxford University Press.

PARFIT, Derek. 2011. *On What Matters*. 2 vols. The Berkeley Tanner Lectures. Nova York: Oxford University Press.

PARRINGTON, Alan J. 1997. “Mutually Assured Destruction Revisited”. *Airpower Journal* 11 (4).

PASQUALOTTO, Emanuele, FEDERICI, Stefano e BELARDINELLI, Marta Olivetti. 2012. “Toward Functioning and Usable Brain-Computer Interfaces (bcis): A Literature Review”. *Disability and Rehabilitation: Assistive Technology* 7 (2): 89-103.

PEARL, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2 ed. Nova York: Cambridge University Press.

PERLMUTTER, J. S. e MINK, J. W. 2006. “Deep Brain Stimulation”. *Annual Review of Neuroscience* 29: 229-57.

PINKER, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. Nova York: Viking.

PLOMIN, R., HAWORTH, C. M., MEABURN, E. L., PRICE, T. S., WELLCOME TRUST CASE CONTROL CONSORTIUM 2 e DAVIS, O. S. 2013. “Common dna Markers Can Account for More than Half of the Genetic Influence on Cognitive Abilities”. *Psychological Science* 24 (2): 562-8.

POPPER, Nathaniel. 2012. “Flood of Errant Trades Is a Black Eye for Wall Street”. *New York Times*, 1 ago.

POURRET, Olivier, NAIM, Patrick e MARCOT, Bruce, eds. 2008. *Bayesian Networks: A Practical Guide to Applications*. Chichester, West Sussex, UK: Wiley.

POWELL, A., SHENNAN, S. e THOMAS, M. G. 2009. “Late Pleistocene Demography and the Appearance of Modern Human Behavior”. *Science* 324 (5932): 1298-1301.

PRICE, Huw. 1991. “Agency and Probabilistic Causality”. *British Journal for the Philosophy of Science* 42 (2): 157-76.

QIAN, M., WANG, D., WATKINS, W. E., GEBSKI, V., YAN, Y. Q., LI, M. e CHEN, Z. P. 2005. “The Effects of Iodine on Intelligence in Children: A Meta-Analysis of Studies Conducted in China”. *Asia Pacific Journal of Clinical Nutrition* 14 (1): 32-42.

QUINE, Willard Van Orman e ULLIAN, Joseph Silbert. 1978. *The Web of Belief,* ed. Richard Malin Ohmann, vol. 2. Nova York: Random House.

RAILTON, Peter. 1986. “Facts and Values”. *Philosophical Topics* 14 (2): 5-31.

RAJAB, Moheeb Abu, ZARFOSS, Jay, MONROSE, Fabian e TERZIS, Andreas. 2006. “A Multifaceted Approach to Understanding the Botnet Phenomenon”. In *Proceedings of the 6th acm sigcomm Conference on Internet Measurement,* 41-52. Nova York: acm.

RAWLS, John. 1971. *A Theory of Justice*. Cambridge, MA: Belknap.

READ, J. I. e TRENTHAM, Neil. 2005. “The Baryonic Mass Function of Galaxies”. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363 (1837): 2693-710.

REPANTIS, D., SCHLATTMANN, P., LAISNEY, O. e HEUSER, I. 2010. “Modafinil and Methylphenidate for Neuroenhancement in Healthy Individuals: A Systematic Review”. *Pharmacological Research* 62 (3): 187-206.

RHODES, Richard. 1986. *The Making of the Atomic Bomb*. Nova York: Simon & Schuster.

RHODES, Richard. 2008. *Arsenals of Folly: The Making of the Nuclear Arms Race*. Nova York: Vintage.

RIETVELD, Cornelius A., MEDLAND, Sarah E., DERRINGER, Jaime, YANG, Jian, ESKO, Tonu, MARTIN, Nicolas W., WESTRA, Harm-Jan, SHAKHBAZOV, Konstantin, ABDELLAOUI, Abdel, et al. 2013. “gwas of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment”. *Science* 340 (6139): 1467-71.

RING, Mark e ORSEAU, Laurent. 2011. “Delusion, Survival, and Intelligent Agents”. In *Artificial General Intelligence: 4th International Conference, agi 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, editado por Jürgen Schmidhuber, Kristinn R. Thórisson e Moshe Looks, 11-20. Lecture Notes in Computer Science 6830. Berlim: Springer.

RITCHIE, Graeme, MANURUNG, Ruli e WALLER, Annalu. 2007. “A Practical Application of Computational Humour”. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, editado por Amilcar Cardoso e GERAINT A. Wiggins, 91-8. Londres: Goldsmiths, University of London.

ROACHE, Rebecca. 2008. “Ethics, Speculation, and Values”. *NanoEthics* 2 (3): 317-27.

ROBLES, J. A., LINEWEAVER, C. H., GRETHER, D., FLYNN, C., EGAN, C. A., PRACY, M. B., HOLMBERG, J. e GARDNER, E. 2008. “A Comprehensive Comparison of the Sun to Other Stars: Searching for Self-Selection Effects”. *Astrophysical Journal* 684 (1): 691-706.

ROE, Anne. 1953. *The Making of a Scientist*. Nova York: Dodd, Mead.

ROY, Deb. 2012. “About”. Recuperado em 14 out. Disponível em http://web.media.mit.edu/~dkroy/.

RUBIN, Jonathan e WATSON, Ian. 2011. “Computer Poker: A Review”. *Artificial Intelligence* 175 (5-6): 958-87.

RUMELHART, D. E., HINTON, G. E. e WILLIAMS, R. J. 1986. “Learning Representations by Back-Propagating Errors”. *Nature* 323 (6088): 533-6.

RUSSELL, Bertrand. 1986. “The Philosophy of Logical Atomism”. In *The Philosophy of Logical Atomism and Other Essays 1914-1919*, editado por John G. Slater, 8: 157-244. The Collected Papers of Bertrand Russell. Boston: Allen & Unwin.

RUSSELL, Bertrand e GRIFFIN, Nicholas. 2001. *The Selected Letters of Bertrand Russell: The Public Years, 1914-1970*. Nova York: Routledge.

RUSSELL, Stuart J. e NORVIG, Peter. 2010. *Artificial Intelligence: A Modern Approach*. 3 ed. Upper Saddle River, NJ: Prentice-Hall.

SABROSKY, Curtis W. 1952. “How Many Insects Are There?”. In *Insects*, editado por United States Department of Agriculture, 1-7. Yearbook of Agriculture. Washington, DC: United States Government Printing Office.

SALAMON, Anna. 2009. “When Software Goes Mental: Why Artificial Minds Mean Fast Endogenous Growth”. Documento de trabalho, 27 dez.

SALEM, D. J. e ROWAN, A. N. 2001. *The State of the Animals: 2001*. Public Policy Series. Washington, DC: Humane Society Press.

SALVERDA, W., NOLAN, B. e SMEEDING, T. M. 2009. *The Oxford Handbook of Economic Inequality*. Oxford: Oxford University Press.

SAMUEL, A. L. 1959. “Some Studies in Machine Learning Using the Game of Checkers”. *IBM Journal of Research and Development* 3 (3): 210-19.

SANDBERG, Anders. 1999. “The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains”. *Journal of Evolution and Technology* 5.

SANDBERG, Anders. 2010. “An Overview of Models of Technological Singularity”. Artigo apresentado no workshop Roadmaps to agi and the Future of agi, Lugano, Suíça, 8 mar.

SANDBERG, Anders. 2013. “Feasibility of Whole Brain Emulation”. In *Philosophy and Theory of Artificial Intelligence*, editado por Vincent C. Müller, 5: 251-64. Studies in Applied Philosophy, Epistemology and Rational Ethics. Nova York: Springer.

SANDBERG, Anders e BOSTROM, Nick. 2006. “Converging Cognitive Enhancements”. *Annals of the Nova York Academy of Sciences* 1093: 201-27.

SANDBERG, Anders e BOSTROM, Nick. 2008. *Whole Brain Emulation: A Roadmap*. Relatório técnico 2008-3. Future of Humanity Institute, University of Oxford.

SANDBERG, Anders e BOSTROM, Nick. 2011. *Machine Intelligence Survey*. Relatório técnico 2011-1. Future of Humanity Institute, University of Oxford.

SANDBERG, Anders e Savulescu, Julian. 2011. “The Social and Economic Impacts of Cognitive Enhancement”. In *Enhancing Human Capacities*, editado por Julian Savulescu, Ruud ter Meulen e Guy Kahane, 92-112. Malden, MA: Wiley-Blackwell.
SCHAEFFER, Jonathan. 1997. *One Jump Ahead: Challenging Human Supremacy in Checkers*. Nova York: Springer.
SCHAEFFER, J., BURCH, N., BJORNSSON, Y., KISHIMOTO, A., MULLER, M., LAKE, R., Lu, P. e Sutphen, S. 2007. “Checkers Is Solved”. *Science* 317 (5844): 1518-22.
SCHALK, Gerwin. 2008. “Brain-Computer Symbiosis”. *Journal of Neural Engineering* 5 (1): P1-P15.
SCHELLING, Thomas C. 1980. *The Strategy of Conflict*. 2 ed. Cambridge, MA: Harvard University Press.
SCHULTZ, T. R. 2000. “In Search of Ant Ancestors”. *Proceedings of the National Academy of Sciences of the United States of America* 97 (26): 14028-9.
SCHULTZ, W., Dayan, P. e Montague, P. R. 1997. “A Neural Substrate of Prediction and Reward”. *Science* 275 (5306): 1593-9.
SCHWARTZ, Jacob T. 1987. “Limits of Artificial Intelligence”. In *Encyclopedia of Artificial Intelligence*, editado por Stuart C. Shapiro e David Eckroth, 1: 488-503. Nova York: Wiley.
SCHWITZGEBEL, Eric. 2013. “If Materialism is True, the United States is Probably Conscious”. Documento de trabalho, 8 fev.
SEN, Amartya e WILLIAMS, Bernard, eds. 1982. *Utilitarianism and Beyond*. Nova York: Cambridge University Press.
SHANAHAN, Murray. 2010. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. Nova York: Oxford University Press.
SHANNON, Robert V. 2012. “Advances in Auditory Prostheses”. *Current Opinion in Neurology* 25 (1): 61-6.
SHAPIRO, Stuart C. 1992. “Artificial Intelligence”. In *Encyclopedia of Artificial Intelligence*, 2 ed., 1: 54-7. Nova York: Wiley.
SHEPPARD, Brian. 2002. “World-Championship-Caliber Scrabble”. *Artificial Intelligence* 134 (1-2): 241-75.
SHOEMAKER, Sydney. 1969. “Time Without Change”. *Journal of Philosophy* 66 (12): 363-81.
SHULMAN, Carl. 2010a. *Omohundro’s “Basic AI Drives” and Catastrophic Risks*. Berkeley, CA: Machine Intelligence Research Institute.
SHULMAN, Carl. 2010b. *Whole Brain Emulation and the Evolution of Superorganisms.* Berkeley, CA: Machine Intelligence Research Institute.
SHULMAN, Carl. 2012. “Could We Use Untrustworthy Human Brain Emulations to Make Trustworthy Ones?”. Artigo apresentado na conferência agi Impacts, Oxford, 8-11 dez.
SHULMAN, Carl e BOSTROM, Nick. 2012. “How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects”. *Journal of Consciousness Studies* 19 (7-8): 103-30.
SHULMAN, Carl e BOSTROM, Nick. 2014. “Embryo Selection for Cognitive Enhancement: Curiosity or Game-Changer?”. *Global Policy* 5 (1): 85-92.
SHULMAN, Carl, JONSSON, Henrik e TARLETON, Nick. 2009. “Which Consequentialism? Machine Ethics and Moral Divergence”. In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan*. *Proceedings*, editado por Carson Reynolds e Alvaro Cassinelli, 23-25. AP-CAP 2009.
SIDGWICK, Henry e JONES, Emily Elizabeth Constance. 2010. *The Methods of Ethics*. Charleston, SC: Nabu Press.
SILVER, Albert. 2006. “How Strong Is gnu Backgammon?”. Backgammon Galore! 16 set. Recuperado em 26 out. 2013. Disponível em www.bkgm.com/gnu/AllAboutGNU.html#how_strong_is_gnu.
SIMERAL, J. D., KIM, S. P., BLACK, M. J., DONOGHUE, J. P. e HOCHBERG, L. R. 2011. “Neural Control of Cursor Trajectory and Click by a Human with Tetraplegia 1000 Days after Implant of an Intracortical Microelectrode Array”. *Journal of Neural Engineering* 8 (2): 025027.
SIMESTER, Duncan e KNEZ, Marc. 2002. “Direct and Indirect Bargaining Costs and the Scope of the Firm”. *Journal of Business* 75 (2): 283-304.
SIMON, Herbert Alexander. 1965. *The Shape of Automation for Men and Management*. Nova York: Harper & Row.
SINHABABU, Neil. 2009. “The Humean Theory of Motivation Reformulated and Defended”. *Philosophical Review* 118 (4): 465-500.
SLAGLE, James R. 1963. “A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus”. *Journal of the acm* 10 (4): 507-20.
SMEDING, H. M., SPEELMAN, J. D., KONING-HAANSTRA, M., SCHUURMAN, P. R., NIJSSEN, P., VAN LAAR, T. e SCHMAND, B. 2006. “Neuropsychological Effects of Bilateral stn Stimulation in Parkinson Disease: A Controlled Study”. *Neurology* 66 (12): 1830-6.

SMITH, Michael. 1987. “The Humean Theory of Motivation”. *Mind: A Quarterly Review of Philosophy* 96 (381): 36-61.
SMITH, Michael, LEWIS, David e JOHNSTON, Mark. 1989. “Dispositional Theories of Value”. *Proceedings of the Aristotelian Society* 63: 89-174.
SPARROW, Robert. 2013. “In Vitro Eugenics”. *Journal of Medical Ethics*. doi:10.1136/medethics-2012-101200. Publicado on-line em 4 abr. 2013. Disponível em http://jme.bmj.com/content/early/2013/02/13/medethics-2012-101200.full.
STANSBERRY, Matt e KUDRITZKI, Julian. 2012. *Uptime Institute 2012 Data Center Industry Survey*. Uptime Institute.
STAPLEDON, Olaf. 1937. *Star Maker*. Londres: Methuen.
STERIADE, M., TIMOFEEV, I., DURMULLER, N. e GRENIER, F. 1998. “Dynamic Properties of Corticothalamic Neurons and Local Cortical Interneurons Generating Fast Rhythmic (30-40 Hz) Spike Bursts”. *Journal of Neurophysiology* 79 (1): 483-90.
STEWART, P. W., LONKY, E., REIHMAN, J., PAGANO, J., GUMP, B. B. e DARVILL, T. 2008. “The Relationship Between Prenatal pcb Exposure and Intelligence (IQ) in 9-Year-Old Children”. *Environmental Health Perspectives* 116 (10): 1416-22.
SUN, W., YU, H., SHEN, Y., BANNO, Y., XIANG, Z. e ZHANG, Z. 2012. “Phylogeny and Evolutionary History of the Silkworm”. *Science China Life Sciences* 55 (6): 483-96.
SUNDET, J., BARLAUG, D. e TORJUSSEN, T. 2004. “The End of the Flynn Effect? A Study of Secular Trends in Mean Intelligence Scores of Norwegian Conscripts During Half a Century”. *Intelligence* 32 (4): 349-62.
SUTTON, Richard S. e BARTO, Andrew G. 1998. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. Cambridge, MA: mit Press.
TALUKDAR, D., SUDHIR, K. e AINSLIE, A. 2002. “Investigating New Product Diffusion Across Products and Countries”. *Marketing Science* 21 (1): 97-114.
TEASDALE, Thomas W. e OWEN, David R. 2008. “Secular Declines in Cognitive Test Scores: A Reversal of the Flynn Effect”. *Intelligence* 36 (2): 121-6.
TEGMARK, Max e BOSTROM, Nick. 2005. “Is a Doomsday Catastrophe Likely?”. *Nature* 438: 754.
TEITELMAN, Warren. 1966. “Pilot: A Step Towards Man-Computer Symbiosis”. Dissertação de doutorado, Massachusetts Institute of Technology.
TEMPLE, Robert K. G. 1986. *The Genius of China: 3000 Years of Science, Discovery, and Invention*. 1 ed. Nova York: Simon & Schuster.
TESAURO, Gerald. 1995. “Temporal Difference Learning and TD-Gammon”. *Communications of the acm* 38 (3): 58-68.
TETLOCK, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton, NJ: Princeton University Press.
TETLOCK, Philip E. e BELKIN, Aaron. 1996. “Counterfactual Thought Experiments in World Politics: Logical, Methodological e Psychological Perspectives”. In *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, editado por Philip E. Tetlock e Aaron Belkin, 1-38. Princeton, NJ: Princeton University Press.
THOMPSON, Adrian. 1997. “Artificial Evolution in the Physical World”. In *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, editado por Takashi Gomi, 101-25. ER ’97. Carp, ON: Applied AI Systems.
THRUN, S., MONTEMERLO, M., DAHLKAMP, H., STAVENS, D., ARON, A., DIEBEL, J., FONG, P., et al. 2006. “Stanley: The Robot That Won the darpa Grand Challenge”. *Journal of Field Robotics* 23 (9): 661-92.
TRACHTENBERG, J. T., CHEN, B. E., KNOTT, G. W., FENG, G., SANES, J. R., WELKER, E. e SVOBODA, K. 2002. “Long-Term In Vivo Imaging of Experience-Dependent Synaptic Plasticity in Adult Cortex”. *Nature* 420 (6917): 788-94.
TRAUB, Wesley A. 2012. “Terrestrial, Habitable-Zone Exoplanet Frequency from *Kepler*”. *Astrophysical Journal* 745 (1): 1-10.
TRUMAN, James W., TAYLOR, Barbara J. e AWAD, Timothy A. 1993. “Formation of the Adult Nervous System”. In *The Development of* Drosophila Melanogaster, editado por Michael Bate e Alfonso Martinez Arias. Plainview, NY: Cold Spring Harbor Laboratory.
TUOMI, Ilkka. 2002. “The Lives and the Death of Moore’s Law”. *First Monday* 7 (11).
TURING, A. M. 1950. “Computing Machinery and Intelligence”. *Mind* 59 (236): 433-60.
TURKHEIMER, Eric, HALEY, Andreana, WALDRON, Mary, D’ONOFRIO, Brian e GOTTESMAN, Irving I. 2003. “Socioeconomic Status Modifies Heritability of IQ in Young Children”. *Psychological Science* 14 (6): 623-8.
UAUY, Ricardo e DANGOUR, Alan D. 2006. “Nutrition in Brain Development and Aging: Role of Essential Fatty Acids”. Suplemento, *Nutrition Reviews* 64 (5): S24-S33.
ULAM, Stanislaw M. 1958. “John von Neumann”. *Bulletin of the American Mathematical Society* 64 (3): 1-49.

Uncertain Future, The. 2012. “Frequently Asked Questions”. The Uncertain Future. Recuperado em 25 mar. 2012. Disponível em www.theuncertainfuture.com/faq.html.
U.S. CONGRESS, Office of Technology Assessment. 1995. *U.S.-Russian Cooperation in Space* ISS-618. Washington, DC: U.S. Government Printing Office, abr.
VAN ZANDEN, Jan Luiten. 2003. *On Global Economic History: A Personal View on an Agenda for Future Research*. International Institute for Social History, 23 jul.
VARDI, Moshe Y. 2012. “Artificial Intelligence: Past and Future”. *Communications of the acm* 55 (1): 5.
VASSAR, Michael e FREITAS, Robert A., Jr. 2006. “Lifeboat Foundation Nanoshield”. Lifeboat Foundation. Recuperado em 12 maio 2012. Disponível em http://lifeboat.com/ex/nanoshield.
VINGE, Vernor. 1993. “The Coming Technological Singularity: How to Survive in the Post-Human Era”. In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, 11-22. nasa Conference Publication 10129. nasa Lewis Research Center.
VISSCHER, P. M., HILL, W. G. e WRAY, N. R. 2008. “Heritability in the Genomics Era: Concepts and Misconceptions”. *Nature Reviews Genetics* 9 (4): 255-66.
VOLLENWEIDER, Franz, GAMMA, Alex, LIECHTI, Matthias e HUBER, Theo. 1998. “Psychological and Cardiovascular Effects and Short-Term Sequelae of mdma (‘Ecstasy’) in mdma-Naïve Healthy Volunteers”. *Neuropsychopharmachology* 19 (4): 241-51.
WADE, Michael J. 1976. “Group Selections Among Laboratory Populations of Tribolium”. *Proceedings of the National Academy of Sciences of the United States of America* 73 (12): 4604-7.
WAINWRIGHT, Martin J. e JORDAN, Michael I. 2008. “Graphical Models, Exponential Families, and Variational Inference”. *Foundations and Trends in Machine Learning* 1 (1-2): 1-305.
WALKER, Mark. 2002. “Prolegomena to Any Future Philosophy”. *Journal of Evolution and Technology* 10 (1).
WALSH, Nick Paton. 2001. “Alter our dna or robots will take over, warns Hawking”. *The Observer*, 1 set. Disponível em www.theguardian.com/uk/2001/sep/02/medicalscience.genetics.
WARWICK, Kevin. 2002. *I, Cyborg*. Londres: Century.
WEHNER, M., OLIKER, L. e SHALF, J. 2008. “Towards Ultra-High Resolution Models of Climate and Weather”. *International Journal of High Performance Computing Applications* 22 (2): 149-65.
WEIZENBAUM, Joseph. 1966. “Eliza: A Computer Program for the Study of Natural Language Communication Between Man And Machine”. *Communications of the acm* 9 (1): 36-45.
WEIZENBAUM, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. São Francisco, CA: W. H. Freeman.
WERBOS, Paul John. 1994. *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*. Nova York: Wiley.
WHITE, J. G., SOUTHGATE, E., THOMSON, J. N. e BRENNER, S. 1986. “The Structure of the Nervous System of the Nematode *Caenorhabditis Elegans*”. *Philosophical Transactions of the Royal Society of London*. *Series B, Biological Sciences* 314 (1165): 1-340.
WHITEHEAD, Hal. 2003. *Sperm Whales: Social Evolution in the Ocean*. Chicago: University of Chicago Press.
WHITMAN, William B., COLEMAN, David C. e WIEBE, William J. 1998. “Prokaryotes: The Unseen Majority”. *Proceedings of the National Academy of Sciences of the United States of America* 95 (12): 6578-83.
WIENER, Norbert. 1960. “Some Moral and Technical Consequences of Automation”. *Science* 131 (3410): 1355-8.
WIKIPEDIA. 2012a, s.v. “Computer Bridge”. Recuperado em 30 jun. 2013. Disponível em http://en.wikipedia.org/wiki/Computer_bridge.
WIKIPEDIA. 2012b, s.v. “Supercomputer”. Recuperado em 30 jun. 2013. Disponível em http://et.wikipedia.org/wiki/Superarvuti.
WILLIAMS, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton Science Library. Princeton, NJ: Princeton University Press.
WINOGRAD, Terry. 1972. *Understanding Natural Language*. Nova York: Academic Press.
WOOD, Nigel. 2007. *Chinese Glazes: Their Origins, Chemistry and Re-creation*. Londres: A&C Black.
WORLD BANK. 2008. *Global Economic Prospects: Technology Diffusion in the Developing World*, 42097. Washington, DC.
WORLD ROBOTICS. 2011. *Executive Summary of 1. World Robotics 2011 Industrial Robots; 2. World Robotics 2011 Service Robots.* Recuperado em 30 jun. 2012. Disponível em www.bara.org.uk/pdf/2012/world-robotics/Executive_Summary_WR_2012.pdf.
WORLD VALUES SURVEY. 2008. *wvs 2005-2008*. Recuperado em 29 out. 2013. Disponível em www.wvsevsdb.com/wvs/WVSAnalizeStudy.jsp.
WRIGHT, Robert. 2001. *Nonzero: The Logic of Human Destiny*. Nova York: Vintage.
YAEGER, Larry. 1994. “Computational Genetics, Physiology, Metabolism, Neural Systems, Learning, Vision, and Behavior or PolyWorld: Life in a New Context”. In *Proceedings of the Artificial Life III Conference*, editado por C. G. Langton, 263-98. Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley.

YUDKOWSKY, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* Berkeley, CA: Machine Intelligence Research Institute, 15 jun.
YUDKOWSKY, Eliezer. 2002. “The AI-Box Experiment”. Recuperado em 15 jan. 2012. Disponível em http://yudkowsky.net/singularity/aibox.
YUDKOWSKY, Eliezer. 2004. *Coherent Extrapolated Volition.* Berkeley, CA: Machine Intelligence Research Institute, maio.
YUDKOWSKY, Eliezer. 2007. “Levels of Organization in General Intelligence”. In *Artificial General Intelligence*, editado por Ben Goertzel e Cassio Pennachin, 389-501. Cognitive Technologies. Berlim: Springer.
YUDKOWSKY, Eliezer. 2008a. “Artificial Intelligence as a Positive and Negative Factor in Global Risk”. In *Global Catastrophic Risks*, editado por Nick Bostrom e Milan M. Ćirković, 308-45. Nova York: Oxford University Press.
YUDKOWSKY, Eliezer. 2008b. “Sustained Strong Recursion”. *Less Wrong* (blog), 5 dez.
YUDKOWSKY, Eliezer. 2010. *Timeless Decision Theory*. Berkeley, CA: Machine Intelligence Research Institute.
YUDKOWSKY, Eliezer. 2011. *Complex Value Systems are Required to Realize Valuable Futures*. Berkeley, CA: Machine Intelligence Research Institute.
YUDKOWSKY, Eliezer. 2013. *Intelligence Explosion Microeconomics*, Relatório técnico 2013-1. Berkeley, CA: Machine Intelligence Research Institute.
ZAHAVI, Amotz e Zahavi, Avishag. 1997. *The Handicap Principle: A Missing Piece of Darwin’s Puzzle.* Trad. N. Zahavi-Ely e M. P. Ely. Nova York: Oxford University Press.
ZALASIEWICZ, J., WILLIAMS, M., SMITH, A., BARRY, T. L., COE, A. L., BOWN, P. R., BRENCHLEY, P., et al. 2008. “Are We Now Living in the Anthropocene?”. *GSA Today* 18 (2): 4-8.
ZEIRA, Joseph. 2011. “Innovations, Patent Races and Endogenous Growth”. *Journal of Economic Growth* 16 (2): 135-56.
ZULETA, Hernando. 2008. “An Empirical Note on Factor Shares”. *Journal of International Trade and Economic Development* 17 (3): 379-90.

AGRADECIMENTOS

A membrana que envolve o processo de escrita tem sido bastante permeável. Assim como muitos conceitos e ideias gerados durante a escrita deste livro “vazaram” e se tornaram parte de uma discussão mais ampla, vários insights vindos de fora enquanto o livro estava sendo escrito foram incorporados ao texto. Tentei ser diligente com os instrumentos de citação, mas as influências são tantas que se torna difícil registrar todas elas.

Pelas discussões mais amplas que colaboraram para elucidar meu pensamento, devo meus agradecimentos a um grupo considerável de pessoas, incluindo Ross Andersen, Stuart Armstrong, Owen Cotton-Barrat, Nick Beckstead, Yoshua Bengio, David Chalmers, Paul Christiano, Milan Ćirković, Andrew Critch, Daniel Dennett, David Deutsch, Daniel Dewey, Thomas Dietterich, Eric Drexler, David Duvenaud, Peter Eckersley, Amnon Eden, Oren Etzioni, Owain Evans, Benja Fallenstein, Alex Flint, Carl Frey, Zoubin Ghahramani, Ian Goldin, Katja Grace, Roger Grosse, Tom Gunter, J. Storrs Hall, Robin Hanson, Demis Hassabis, Geoffrey Hinton, James Hughes, Marcus Hutter, Garry Kasparov, Marcin Kulczycki, Patrick LaVictoire, Shane Legg, Moshe Looks, Willam MacAskill, Eric Mandelbaum, Gary Marcus, James Martin, Lillian Martin, Roko Mijic, Vincent Mueller, Elon Musk, Séan Ó hÉigeartaigh, Christopher Olah, Toby Ord, Laurent Orseau, Michael Osborne, Larry Page, Dennis Pamlin, Derek Parfit, David Pearce, Huw Price, Guy Ravine, Martin Rees, Bill Roscoe, Francesca Rossi, Stuart Russell, Anna Salamon, Lou Salkind, Anders Sandberg, Julian Savulescu, Jürgen Schmidhuber, Bart Selman, Nicholas Shackel, Murray Shanahan, Noel Sharkey, Carl Shulman, Peter Singer, Nate Soares, Dan Stoicescu, Mustafa Suleyman, Jaan Tallinn, Alexander Tamas, Jessica Taylor, Max Tegmark, Roman Yampolskiy e Eliezer Yudkowsky.

Pelos comentários especialmente detalhados, sou grato a Milan Ćirković, Daniel Dewey, Owain Evans, Nick Hay, Keith Mansfield, Luke Muehlhauser, Toby Ord, Jess Riedel, Anders Sandberg, Murray Shanahan e Carl Shulman. Pelo aconselhamento e auxílio na pesquisa para diferentes partes do livro, gostaria de agradecer Stuart Armstrong, Daniel Dewey, Eric Drexler, Alexandre Erler, Rebecca Roache e Anders Sandberg.

Pela ajuda na preparação dos manuscritos, agradeço a Caleb Bell, Malo Bourgon, Robin Brandt, Lance Bush, Cathy Douglass, Alexandre Erler, John King, Kristian Rönn, Susan Rogers, Kyle Scott, Andrew Snyder-Beattie, Cecilia Tilli e Alex Vermeer. Gostaria ainda de agradecer especialmente ao meu editor, Keith Mansfield, pelo encorajamento incessante ao longo de todo o projeto.

Peço desculpas a todos os outros que deveriam ter sido lembrados aqui.

E, finalmente, meu mais profundo agradecimento aos financiadores, amigos e à família: sem o apoio de vocês, esse trabalho não teria sido possível.

Nick Bostrom é professor na Faculdade de Filosofia na Universidade de Oxford, no Reino Unido, e diretor fundador do Instituto para o Futuro da Humanidade e do Programa sobre os Impactos da Tecnologia do Futuro, na Oxford Martin School. É autor de mais de duzentas publicações, entre elas, *Anthropic Bias* (2002), *Global Catastrophic Risks* (2008) e *Human Enhancement* (2009). Foi professor em Yale e Postdoctoral Fellow da British Academy. Tem formação em física, neurociência computacional, lógica matemática e filosofia. Saiba mais em nickbostrom.com.

CRÂNIO — a nova linha editorial de não ficção da DarkSide® Books — estimula o leitor a entender e questionar o mundo que estamos construindo. Após desenterrar clássicos inesquecíveis e revelar novos fenômenos da literatura dark, a 1ª editora brasileira inteiramente dedicada ao terror e à fantasia amplia seus horizontes. O objetivo é trilhar novos caminhos, mostrando que ciência, inovação, história e filosofia podem ser tão surpreendentes quanto a mais criativa obra de ficção.

Aqui tudo é real. E ainda assim, fantástico e muitas vezes assustador. Assuntos delicados e surpreendentes são tratados com o respeito que merecem, com uma linguagem que aproxima o leitor. Devorar um título da série CRÂNIO é aceitar um convite à reflexão do agora.

O compromisso da linha editorial CRÂNIO é publicar material minuciosamente selecionado. Livros assinados por especialistas, acadêmicos e pensadores em diversas áreas, dispostos a dividir experiências e pontos de vista transformadores que nos ajudem a entender melhor esse estranho e admirável mundo novo.



Copyright © Nick Bostrom 2014

Superintelligence: Paths, Dangers, Strategies was originally published in English in 2014. This translation is published by arrangement with Oxford University Press. DarkSide® Books is solely responsible for this translation from the original work and Oxford University Press shall have no liability for any errors, omissions or inaccuries or ambiguities in such translation or for any losses caused by reliance there on.

Superinteligência: caminhos, perigos, estratégias foi publicado originalmente em inglês em 2014. Esta tradução é publicada mediante acordo com Oxford University Press. A DarkSide® Books é a única responsável por esta tradução da obra original, e a Oxford University Press não se responsabiliza por quaisquer erros, omissões, imprecisões ou ambiguidades nessa tradução, ou por quaisquer perdas causadas pela mesma.

Todos os direitos reservados.

Ilustração da capa e p. 10 © Claire Scully
Design da capa © Nick Bostrom

Tradução para a língua portuguesa
© Aurélio Antônio Monteiro, Clemente Gentil Penna, Fabiana Geremias Monteiro, Patrícia Ramos Geremias, 2018

[2018]
Todos os direitos desta edição reservados à
DarkSide® Entretenimento LTDA.
Rua do Russel, 450/501 — 22210-010
Glória — Rio de Janeiro — RJ — Brasil
www.darksidebooks.com

Diretor Editorial
Christiano Menezes

Diretor Comercial
Chico de Assis

Gerente de Novos Negócios
Frederico Nicolay

Gerente de Marketing Digital
Mike Ribera

Editores
Bruno Dorigatti
Raquel Moritz

Editores Assistentes
Lielson Zeni
Nilsen Silva

Design e Adaptação da Capa
Retina 78

Designers Assistentes
Marco Luz
Pauline Qui

Revisão Técnica
Bianca Zadrozny, Ph.D.

Revisão
Cecilia Floresta
Felipe Pontes

Podução de ebook
S2 Books

Isbn: 978-85-9454-073-7



[1] Entretanto, nem todas as notas têm informações úteis.

[2] Não sei quais deles.

[3] A renda em nível de subsistência está estimada atualmente em US$ 400 (Chen e Ravallion, 2010). Um milhão de rendas de subsistência é, então, US$ 400.000.000. O produto bruto mundial atualmente é de aproximadamente US$ 60.000.000.000.000 e em anos recentes têm crescido numa taxa anual de aproximadamente 4% (taxa de crescimento anual composto desde 1950, com base em Maddison [2010]). Esses números produzem a estimativa mencionada no texto, que é, obviamente, apenas uma aproximação da ordem de magnitude. Se olharmos diretamente para as taxas populacionais, veremos que atualmente a população mundial leva em média uma semana e meia para crescer em 1 milhão; mas isso subestima a taxa de crescimento da economia, já que a renda per capita também está aumentando. Em 5000 a.C., após a revolução agrícola, a população mundial estava crescendo na proporção de 1 milhão a cada duzentos anos — uma grande aceleração, se considerarmos a proporção de, talvez, 1 milhão a cada 1 milhão de anos no início da pré-história da humanidade; ou seja, uma grande aceleração já havia ocorrido até então. Mesmo assim, é impressionante que um volume de crescimento econômico que há 7 mil anos levava duzentos anos leve, atualmente, apenas noventa minutos para ocorrer, e que um aumento na população mundial que levou dois séculos para acontecer no passado demore apenas uma semana e meia nos tempos de agora. Ver também Maddison (2005).

[4] Esse crescimento e aceleração acentuados podem sugerir o surgimento de uma possível “singularidade”, como descrito por John von Neumann numa conversa com o matemático Stanislaw Ulam: “Nossa conversa se centrou no progresso cada vez mais acelerado da tecnologia e nas transformações no modo de vida humana, os quais dão a impressão de que estamos nos aproximando de uma certa singularidade essencial na história da raça para além da qual as relações humanas, tais como as conhecemos, poderiam não continuar”. (Ulam, 1958)

[5] Hanson (2000).

[6] Vinge (1993); Kurzweil (2005).

[7] Sandberg (2010).

[8] Duas afirmações otimistas frequentemente repetidas nos anos 1960: “Em vinte anos, as máquinas serão capazes de fazer qualquer trabalho que um homem é capaz de fazer” (Simon, 1965, p. 96); “No tempo de uma geração […] o problema da criação da inteligência artificial será substancialmente resolvido” (Minsky, 1967, p. 2). Para uma revisão sistemática das previsões sobre IA, ver Armstrong e Sotala (2012).

[9] Ver, por exemplo, Baum et al. (2001) e Armstrong e Sotala (2012).

[10] Essa questão pode sugerir, no entanto, que os pesquisadores de IA saibam menos sobre estimativas de prazos de desenvolvimento do que eles supõem — mas isso poderia fazer com que eles tanto superestimem quanto subestimem o tempo para a IA.

[11] Good (1965, p. 33).

[12] Uma exceção é Norbert Wiener, que tem algumas dúvidas sobre as possíveis consequências. Ele escreveu, em 1960: “Se usarmos, para alcançar nossos objetivos, um ente mecânico sobre cuja operação não possamos interferir eficientemente após seu início, pois sua ação seria tão rápida e irreversível que não teríamos os dados necessários para intervir antes da operação estar completa, é bom, então, que tenhamos certeza de que o propósito inserido na máquina seja o propósito que nós realmente desejamos, e não uma mera imitação embelezada” (Wiener, 1960). Ed Fredkin expressou suas preocupações a respeito de uma IA superinteligente numa entrevista descrita em McCorduck (1979). Em 1970, o próprio Good escreveu a respeito dos riscos e chegou mesmo a sugerir a criação de uma associação para lidar com esses perigos (Good [1970]; veja também o seu último artigo [Good, 1982], no qual ele antecipa algumas das ideias da “normatividade indireta” que discutiremos no capítulo 13). Mais tarde, em 1984, Marvin Minsky também escreveu a respeito de muitas das preocupações fundamentais (Minsky, 1984).

[13] Cf. Yudkowsky (2008a). Sobre a importância de avaliar as implicações éticas de tecnologias futuras potencialmente perigosas *antes* que elas se tornem viáveis, ver Roache (2008).

[14] McCorduck (1979).

[15] Newell et al. (1959).

[16] O programa saints, o programa analogy e o programa student, respectivamente. Ver Slagle (1963), Evans (1964, 1968) e Bobrow (1968).

[17] Nilsson (1984).

[18] Weizenbaum (1966).

[19] Winograd (1972).

[20] Cope (1996); Weizenbaum (1976); Moravec (1980); Thrun et al. (2006); Buehler et al. (2009); Koza et al. (2003). O Departamento de Veículos Motores do Estado de Nevada, Estados Unidos, emitiu, em maio de 2012, a primeira licença para um carro não tripulado.

[21] O sistema standup (Ritchie et al., 2007).

[22] Schwartz (1987). Aqui, Schwartz caracteriza uma visão cética que ele pensou estar representada nos escritos de Hubert Dreyfus.

[23] Um crítico incansável durante esse período foi Hubert Dreyfus. Outros céticos importantes da época foram John Lucas, Roger Penrose e John Searle. No entanto, entre estes apenas Dreyfus estava realmente preocupado em refutar afirmações a respeito de quais conquistas práticas nós poderíamos esperar dos paradigmas existentes de IA (apesar de ele parecer aberto à possibilidade de que novos paradigmas pudessem ir mais longe). Os alvos de Searle eram as teorias funcionalistas da filosofia da mente, e não os poderes instrumentais dos sistemas de IA. Lucas e Penrose negaram que um computador tradicional pudesse algum dia ser programado para fazer tudo o que um matemático humano é capaz de fazer, mas eles não negaram que qualquer função particular poderia, em princípio, ser automatizada ou que a IA seria eventualmente capaz de se tornar instrumentalmente muito poderosa. Cicero destacou que "não existe nada tão absurdo que um filósofo nunca tenha dito" (Cicero, 1923, p. 119); apesar disso, é surpreendentemente difícil encontrar *qualquer* pensador importante que tenha negado a possibilidade da superinteligência de máquina no sentido usado neste livro.

[24] Para muitas aplicações, entretanto, o aprendizado que acontece numa rede neural é pouco diferente do aprendizado que se dá numa regressão linear, uma técnica estatística desenvolvida por Adrien-Marie Legendre e Carl Friedrich Gauss no início do século XIX.

[25] O algoritmo básico foi descrito por Arthur Bryson e Yu-Chi Ho como um método multiestágio de otimização dinâmica em 1969. A aplicação às redes neurais foi sugerida por Paul Werbos em 1974 (Werbos, 1994), mas foi apenas após o trabalho de David Rumelhart, Geoffrey Hinton e Ronald Williams em 1986 (Rumelhart et al., 1986) que o método foi introduzido, gradualmente, ao conhecimento da comunidade em geral.

[26] Já havia sido demonstrado que redes sem camadas ocultas têm sua funcionalidade severamente limitada (Minsky e Papert, 1969).

[27] Por exemplo, MacKay (2003).

[28] Murphy (2012).

[29] Pearl (2009).

[30] Suprimimos vários detalhes técnicos aqui para não sobrecarregar a exposição. Revisitaremos algumas dessas questões ignoradas no capítulo 12.

[31] Um programa p é uma descrição de uma cadeia de caracteres x se p, carregado em uma máquina de Turing universal U (específica), produz x; escrevemos isso como $U(p)=x$. (A cadeia de caracteres x aqui representa um mundo possível.) A complexidade de Kolmogorov de x é então $K(x) := min_p\{l(p) : U(p) = x\}$, onde $l(p)$ é o comprimento de p em bits. A probabilidade "Solomonoff" de x é então definida como $M(x):=\sum_{p:U(p)=x} 2^{-\square(p)}$, onde a soma é definida sobre todos os programas p ("mínimos", ou seja, que não necessariamente param de rodar) para os quais U resulta numa cadeia de caracteres iniciada com x (Hutter, 2005).

[32] O condicionamento bayesiano dada a evidência E gera:

$$P_{\text{posterior}}(w) = P_{\text{prior}}(w|E) = \frac{P_{\text{prior}}(E|w)P_{\text{prior}}(w)}{P_{\text{prior}}(E)},$$

(A probabilidade de uma proposição [como E] é a soma das probabilidades dos mundos possíveis nos quais ela é verdadeira.)

[33] Ou escolhe aleatoriamente uma entre as ações possíveis com as maiores utilidades esperadas, no caso de haver um empate.

[34] De maneira mais concisa, a utilidade prevista de uma ação pode ser escrita como

$$EU(a) = \sum_{w \in W} U(w)P(w|a),$$

onde o somatório é feito sobre todos os mundos possíveis.

[35] Ver, por exemplo, Howson e Urbach (1993); Bernardo e Smith (1994); Russell e Norvig (2010).

[36] Samuel (1959); Schaeffer (1997, cap. 6).

[37] Schaeffer et al. (2007).

[38] Berliner (1980 a, b).

[39] Tesauro (1995).

[40] Tais programas incluem GNU (ver Silver [2006]) e Snowie (ver Gammoned.net [2012]).

[41] O próprio Lenat teve participação no processo de fleet-design. Ele escreveu: “Dessa forma, o crédito final do vencedor deve ser entre 60/40% Lenat/Eurisko, embora o ponto significativo aqui seja que nenhuma parte deveria vencer sozinha” (Lenat 1983, p.80).

[42] Lenat (1982, 1983).

[43] Cirasella e Kopec (2006).

[44] Kasparov (1996, 55).

[45] Newborn (2011).

[46] Keim et al. (1999).

[47] Ver Armstrong (2012).

[48] Sheppard (2002).

[49] Wikipedia (2012 a).

[50] Markoff (2011).

[51] Rubin e Watson (2011).

[52] Elyasaf et al. (2011).

[53] KGS (2012).

[54] Wainwright e Jordan (2008). As áreas de aplicação das redes bayesianas são inúmeras; ver, por exemplo, Pourret et al. (2008).

[55] Pode-se perguntar porque tantos detalhes são dados sobre a IA para jogos aqui, já que esta pode parecer, para alguns, uma área pouco importante de aplicação de IA. A resposta é que os jogos oferecem formas de medidas mais claras do desempenho humano versus o desempenho da IA.

[56] Newell et al. (1958, p. 320).

[57] Atribuído em Vardi (2012).

[58] Em 1976, I.J. Good escreveu: “Um programa de computador de nível grão-mestre nos deixaria a um palmo da [ultrainteligência de máquina]” (Good, 1976). Em 1979, Douglas Hofstadter opinou em sua obra vencedora do prêmio Pulitzer, *Gödel, Escher, Bach*: “Pergunta: Haverá um dia programas de xadrez que conseguirão vencer qualquer pessoa? Especulação: Não. Poderão existir programas capazes de vencer qualquer pessoa no xadrez, mas eles não seriam programas exclusivos de xadrez. Eles serão programas de inteligência geral e serão tão temperamentais quanto pessoas. ‘Você quer jogar xadrez?’ ‘Não, estou entediado com o xadrez. Vamos falar de poesia’”. (Hofstadter [1979] 1999, p. 678).

[59] O algoritmo é de busca minimax com poda alfa-beta, usado com uma função de avaliação heurística específica para estados do tabuleiro. Combinado com uma boa biblioteca de aberturas e fechamentos de jogos e outros truques, ele consegue produzir um mecanismo capaz de jogar xadrez.

[60] No entanto, especialmente com o progresso recente no aprendizado de avaliações heurísticas a partir de jogos simulados, muitos dos algoritmos subjacentes também funcionariam bem para outros jogos.

[61] Nilsson (2009, p. 318). Knuth estava, certamente, fazendo uma observação exagerada. Há muitas “atividades de pensamento” que a IA não teve sucesso em realizar — inventar um novo subcampo de matemática pura, criar qualquer tipo de filosofia, escrever um ótimo romance policial, articular um golpe de estado ou criar um novo e impactante produto de consumo.

[62] Shapiro (1992).

[63] Há quem especule que uma razão pela qual tenha sido difícil alcançar habilidades humanas em percepção, controle motor, senso comum e compreensão da linguagem se deve ao fato de nosso cérebro possuir um *wetware* dedicado a essas funções — estruturas neurais que foram aperfeiçoadas durante a evolução. Em contrapartida, pensamento lógico e habilidades como jogar xadrez não são naturais para nós; dessa forma, talvez tenhamos sido forçados a confiar num conjunto limitado de recursos cognitivos de propósito geral para realizar tais tarefas. Talvez o que o nosso cérebro faz quando nos dedicamos a calcular ou a raciocinar de forma explicitamente lógica é, de certa forma, semelhante ao funcionamento de uma “máquina virtual”, uma simulação

mental lenta e desajeitada de um computador de propósito geral. Alguém pode então dizer (de forma um pouco fantasiosa) que um programa clássico de IA não estaria simulando o pensamento humano, mas que o contrário é mais verdadeiro: um humano que esteja pensando logicamente estaria simulando um programa de IA.

[64] Esse exemplo é controverso: uma opinião minoritária, representada aproximadamente por 20% dos adultos nos Estados Unidos, com números semelhantes encontrados em muitas outras nações desenvolvidas, defende que o Sol gira em torno da Terra (Crabtree, 1999; Dean, 2005).

[65] World Robotics (2011).

[66] Estimado a partir dos dados encontrados em Guizzo (2010).

[67] Holley (2009).

[68] Abordagens estatísticas baseadas em regras híbridas são também usadas, mas compõem atualmente uma pequena parte do cenário.

[69] Cross e Walker (1994); Hedberg (2002).

[70] Com base nas estatísticas do TABB Group, uma empresa que realiza pesquisas sobre mercados de capital sediada em Londres e Nova York (comunicação pessoal).

[71] CFTC e SEC (2010). Para uma perspectiva diferente dos acontecimentos de 6 de maio de 2010, ver CME Group (2010).

[72] Nada no texto deveria ser interpretado como um argumento contra algoritmos de negociação de alta frequência, que podem normalmente ter uma função benéfica ao aumentar a liquidez e a eficiência do mercado.

[73] Um susto menor no mercado ocorreu em 1º de agosto de 2012, devido, em parte, ao fato de que o *circuit breaker* não estava também programado para interromper a negociação caso houvesse mudanças extremas no número de ações sendo negociadas (Popper, 2012). Essa questão adianta outro tema que iremos tratar posteriormente: a dificuldade de antecipar todas as maneiras específicas pelas quais alguma regra em particular, aparentemente plausível, possa estar errada.

[74] Nilsson (2009, p. 319).

[75] Minsky (2006); McCarthy (2007); Beal e Winston (2009).

[76] Peter Norvig, comunicação pessoal. Cursos de aprendizado de máquina são também muito populares, refletindo uma onda de modismo do *big data*, que é, até certo ponto, independente da onda de modismo da IA (inspirada, por exemplo, pelo Google e pelo Netflix Prize).

[77] Armstrong e Sotala (2012).

[78] Müller e Bostrom (2016).

[79] Ver Baum et al. (2011), outra pesquisa citada neste, e Sandberg e Bostrom (2011).

[80] Nilsson (2009).

[81] Isso também está condicionado à não ocorrência de uma catástrofe que afete toda a civilização. A definição de HLMI usada por Nilsson é: “IA capaz de realizar 80% das profissões tão bem ou melhor do que os humanos” (Kruel, 2012).

[82] Entrevistas com aproximadamente 28 profissionais de IA e especialistas em áreas relacionadas (na época de produção deste livro) foram publicadas por Kruel (2011).

[83] Há uma literatura substancial que documenta a baixa confiabilidade das previsões de especialistas em várias áreas e há muitos motivos para pensarmos que muitas das descobertas dessas pesquisas também se aplicam ao campo da inteligência artificial. Especialmente porque quem previu tende a superestimar suas próprias previsões, acreditando que elas sejam mais precisas do que realmente são, e, então, atribuem pouca probabilidade para a possibilidade de que suas hipóteses favoritas estejam erradas (Tetlock, 2005). (Diversos outros vieses também foram documentados; ver, por exemplo, Gilovich et al. [2002]). No entanto, a incerteza é um fato inevitável da condição humana, e muitas das nossas ações inevitavelmente se baseiam em expectativas sobre quais consequências de longo prazo são mais ou menos plausíveis: em outras palavras, em previsões probabilísticas. A recusa em oferecer previsões probabilísticas explícitas não faria com que o problema epistemológico desaparecesse, apenas o esconderia da nossa vista (Bostrom, 2007). Em vez disso, deveríamos responder a evidências de confiança exagerada alargando nossos intervalos de confiança (ou “intervalos críveis”) — ou seja, abandonando nossas crenças — e, de uma maneira geral, devemos lutar da melhor maneira possível contra visões tendenciosas, considerando diferentes perspectivas e buscando a honestidade intelectual. A longo prazo, podemos trabalhar também para desenvolver técnicas, métodos de treinamento e instituições que possam nos ajudar a alcançar uma melhor calibragem. Ver também Armstrong e Sotala (2012).

[84] Essa definição se assemelha àquela presente em Bostrom (2003c) e Bostrom (2006a). Ela também pode ser comparada à de Shane Legg (“A inteligência mede a habilidade de um de agente de alcançar objetivos em muitos tipos de ambientes”) e suas versões mais formais (Legg, 2008). É igualmente muito similar à definição de ultrainteligência de Good que enunciamos no capítulo 1 (“uma máquina capaz de superar todas as atividades intelectuais de qualquer homem, independentemente de quão genial ele seja”).

[85] Por essa mesma razão, não faremos qualquer suposição a respeito da capacidade de uma máquina superinteligente possuir “intencionalidade verdadeira” (segundo Searle, seria possível, mas isso parece ser irrelevante para as questões tratadas neste livro). Também não tomaremos partido quanto à questão do internalismo/externalismo do conteúdo mental que vem sendo debatida na literatura filosófica, tampouco quanto à tese da mente estendida (Clark e Chalmers, 1998).

[86] Turing (1950, p. 456).

[87] Turing (1950, p. 456).

[88] Chalmers (2010); Moravec (1976, 1988, 1998, 1999).

[89] Ver Moravec (1976). Um argumento semelhante é melhor desenvolvido por David Chalmers (2010).

[90] Em Shulman e Bostrom (2012), algumas dessas questões são elaboradas com maiores detalhes.

[91] Legg (2008) apresenta raciocínio semelhante para afirmar que os humanos serão capazes de replicar o processo evolutivo em um período de tempo muito mais reduzido e com recursos computacionais limitados (Legg nota, ao mesmo tempo, que a capacidade computacional necessária para replicar o processo evolutivo sem ajustes ainda está longe do nosso alcance). Baum (2004) argumenta que alguns desenvolvimentos relevantes para o surgimento da IA teriam ocorrido antes, considerando que a própria organização do genoma é uma representação valiosa para a criação de algoritmos evolutivos.

[92] Whitman et al. (1998); Sabrosky (1952).

[93] Schultz (2000).

[94] Menzel e Giurfa (2001, p. 62); Truman et al. (1993).

[95] Sandberg e Bostrom (2008).

[96] Ver Legg (2008) para uma discussão mais detalhada sobre esse assunto e a respeito da promessa de funções ou ambientes que determinem aptidão com base em uma paisagem adaptativa suave de testes de inteligência pura.

[97] Para uma taxonomia e uma discussão mais detalhada dos meios pelos quais os engenheiros talvez sejam capazes de superar a seleção evolutiva natural, ver Bostrom e Sandberg (2009b).

[98] A análise tem se centrado no sistema nervoso de criaturas vivas e não faz referência ao que seria necessário para simular um corpo ou o meio ambiente virtual como parte de uma função de aptidão. É plausível que uma função de aptidão adequada poderia testar a competência de um organismo em particular levando em consideração um número bem menor de operações do que seria necessário para simular todo o processamento neural do cérebro de tal organismo ao longo de todo curso de sua vida. Hoje em dia, os programas de IA geralmente se desenvolvem e operam em ambientes muito abstratos (provadores de teoremas em mundos matemáticos simbólicos, agentes em jogos de competição simples etc.). Um cético pode insistir que um ambiente abstrato seria inadequado para a evolução da inteligência geral e preferiria acreditar que o ambiente virtual deveria se parecer muito com o ambiente biológico no qual nossos ancestrais evoluíram. A criação de um mundo virtual fisicamente realístico exigiria um investimento muito maior de recursos computacionais do que a simulação de um mundo simples artificial ou de um domínio abstrato (enquanto a evolução teve acesso a um mundo real fisicamente realístico “de graça”). No caso limite, a insistência na precisão microfísica aumentaria a necessidade computacional a níveis astronômicos. Entretanto, tal pessimismo extremo é, quase certamente, injustificado. Não parece provável que o ambiente mais adequado para a evolução da inteligência seja aquele que imite a natureza da melhor forma possível. Ao contrário, possivelmente seria mais eficiente o uso de um ambiente artificial para a seleção, um bem distinto daquele de nossos ancestrais e especialmente desenvolvido para promover adaptações que aumentem o tipo de inteligência que procuramos evoluir (por exemplo, o raciocínio abstrato e a habilidade de resolução de problemas gerais em oposição a reações instintivas rápidas ou um sistema visual altamente otimizado).

[99] Para uma visão geral sobre a teoria de observação seletiva, ver Bostrom (2002a). Para sua aplicação específica às questões aqui levantadas, ver Shulman e Bostrom (2012). Para uma pequena introdução a respeito do assunto, ver Bostrom (2008b).

[100] Sutton e Barto (1998, p. 21f); Schultz et al. (1997).

[101] Esse conceito foi introduzido por Eliezer Yudkowsky; ver, por exemplo, Yudkowsky (2007).

[102] Esse é o cenário descrito por Good (1965) e Yudkowsky (2007). Entretanto, também é possível considerar uma alternativa na qual a sequência iterativa possua algumas etapas que não envolvam o aprimoramento da inteligência, mas, sim, a simplificação de arquitetura. Ou seja, em alguns estágios, uma IA embrionária poderia reescrever sua própria arquitetura de modo a tornar melhorias subsequentes mais fáceis de serem alcançadas.

[103] Adequadas para habilitar maneiras instrumentalmente utilitárias de funcionamento cognitivo e de comunicação; mas ainda bastante empobrecidas em relação à interface entre músculos e órgãos sensoriais do corpo de um ser humano normal.

[104] Sandberg (2013).

[105] Ver a seção "*Computer Requirements*" em Sandberg e Bostrom (2008, p. 79-81).

[106] Um nível menor de sucesso poderia ser uma simulação cerebral que tivesse uma microdinâmica biológica sugestiva e mostrasse uma variedade substancial de atividades emergentes típicas da espécie em questão, como sono profundo ou plasticidade cerebral. Tal simulação poderia até funcionar como um ambiente de testes para pesquisas neurocientíficas (apesar de levantar algumas sérias questões éticas), mas não contaria como uma emulação completa do cérebro, a não ser que a simulação fosse suficientemente precisa a ponto de poder realizar uma fração substancial do trabalho intelectual do qual o cérebro simulado também fosse capaz. Como regra, podemos dizer que, para que uma simulação de um cérebro humano se equivalha a uma emulação completa do cérebro, ela teria que ser capaz de expressar pensamentos verbais coerentes ou ter condições de aprender a fazê-lo.

[107] Sandberg e Bostrom (2008).

[108] O primeiro mapeamento foi descrito em Albertson e Thomson (1976) e White et al. (1986). A rede combinada (e com eventuais correções) está disponível no site WormAtlas (www.wormatlas.org).

[109] Para uma revisão das tentativas anteriores de emulação do *C. elegans* e seus resultados, ver Kaufman (2011). Kaufman cita um ambicioso doutorando que atua nessa área, David Dalrymple, o qual afirma: "Com as técnicas optogenéticas, nos encontramos em um ponto no qual não parece descabido afirmar que seremos capazes de 'ler' e 'escrever' dados em qualquer parte do sistema nervoso de um *C. elegans* vivo através do uso de sistemas automatizados de alto desempenho [...]. Acredito que serei capaz de fazer isso daqui a dois ou três anos. Eu ficaria extremamente surpreso se a questão ainda for um problema em aberto no ano de 2020" (Dalrymple, 2011). Modelos cerebrais objetivando o realismo biológico criados manualmente (em vez de automaticamente gerados) já alcançaram algumas funcionalidades básicas; ver, por exemplo, Eliasmith et al. (2012).

[110] Porém, temos que considerar que o *Caenorhabditis elegans* possui algumas propriedades especiais muito convenientes. Por exemplo, o organismo é transparente e o padrão de ligações em seu sistema nervoso é o mesmo para todos os indivíduos da espécie.

[111] Se o produto final for IA neuromórfica, e não a emulação completa do cérebro, os principais insights poderiam, ou não, ser derivados das tentativas de simular cérebros *humanos*. É possível que alguns "truques" corticais fossem descobertos durante o estudo de um cérebro animal (não humano). É provável que seja mais fácil trabalhar com cérebros de alguns animais do que com cérebros humanos, além do que cérebros pequenos iriam requerer menores recursos de escaneamento e modelagem. Além disso, pesquisas em cérebros de animais estariam sujeitas a menos regulamentações. É até mesmo concebível que a primeira máquina com inteligência de nível humano seja criada através da emulação completa do cérebro de algum animal adequado para tanto e, depois disso, seriam encontradas maneiras de aprimorar a mente digital resultante dessa emulação. Assim, a humanidade teria sua réplica a partir de um macaco ou rato de laboratório "incrementado".

[112] Uauy e Dangour (2006); Georgieff (2007); Stewart et al. (2008); Eppig et al. (2010); Cotman e Berchtold (2002).

[113] De acordo com a Organização Mundial da Saúde, em 2007 aproximadamente 2 bilhões de pessoas consumiam quantidades insuficientes de iodo (*The Lancet,* 2008). Deficiência *severa* de iodo limita o desenvolvimento neurológico, causando cretinismo, o que resulta numa diminuição de, em média, 12,5 pontos no QI (Qian et al., 2005). Essa condição pode ser facilmente prevenida com a fortificação do sal de cozinha (Horton et al., 2008).

[114] Bostrom e Sandberg (2009a)

[115] Bostrom e Sandberg (2009b). Em testes que medem a memória de curto prazo, atenção etc., o aumento do desempenho *putativo* derivado de melhorias farmacológicas e nutricionais gira tipicamente em torno de 10% a 20%. Mas, em geral, resta a dúvida se os ganhos medidos são reais, se podem ser sustentados no longo prazo e se são um indicativo de resultados que corresponderiam a melhorias em situações da vida real (Repantis et al., 2010). Por exemplo, em alguns casos, pode ter havido uma compensação por meio da deterioração de alguma dimensão do desempenho que não foi medida nos testes (Bostrom e Sandberg, 2006).

[116] Se existisse uma maneira simples de aprimorar a cognição, era de se esperar que a evolução já tivesse tirado proveito disso. Consequentemente, o tipo de nootrópico cujas investigações teriam potencial de serem mais promissoras seria aquele capaz de aumentar a inteligência de alguma forma que tivesse diminuído a aptidão no processo de seleção natural em meios ambientes ancestrais — por exemplo, aumentar o tamanho do cérebro no nascimento ou incrementar o metabolismo cerebral de glicose. Para uma discussão mais detalhada a respeito (além de outras importantes qualificações), ver Bostrom (2009b).

[117] Os espermatozoides são mais difíceis de ser selecionados, pois, diferente dos embriões, eles são formados por uma única célula — e uma célula precisa necessariamente ser destruída para que ocorra o sequenciamento. Ovócitos também são unicelulares; entretanto, a primeira e a segunda divisão celular são assimétricas e produzem uma célula-filha com pouquíssimo citoplasma, o corpúsculo polar. Uma vez que os corpúsculos polares contêm o mesmo genoma que a célula principal e são redundantes (eventualmente, eles irão se degenerar), é possível realizar uma biópsia neles com um intuito seletivo (Gianaroli, 2000).

[118] Cada uma dessas práticas esteve sujeita a algum tipo de controvérsia ética quando de sua introdução, mas têm sido cada vez mais aceitas. As atitudes em relação à engenharia genética e à seleção de embriões variam significativamente de uma cultura para outra, sugerindo que o desenvolvimento e a aplicação de novas técnicas provavelmente acontecerão mesmo que alguns países inicialmente adotem uma posição de cautela, embora a velocidade com a qual isso ocorrerá será influenciada por questões políticas, morais e religiosas.

[119] Davies et al. (2011); Benyamin et al. (2013); Plomin et al. (2013). Ver também Madris (2011); Hsu (2012).

[120] A herdabilidade no sentido amplo do QI de um adulto frequentemente é estimada entre 0,5 a 0,8, considerando-se indivíduos pertencentes às classes médias de nações desenvolvidas (Bouchard, 2004, p. 148). A herdabilidade no sentido restrito, que mede a proporção da variância que pode ser atribuída a fatores genéticos aditivos é menor (em média, entre 0,3 a 0,5), mas ainda substancial (Devlin et al., 1997; Davies et al., 2011; Visscher et al., 2008). Essas estimativas podem mudar em populações e meio ambientes diferentes, já que a herdabilidade varia de acordo com a população ou o meio ambiente em estudo. Por exemplo, uma herdabilidade menor pode ser percebida entre crianças e indivíduos em ambientes pouco desenvolvidos (Benyamin et al., 2013; Turkheimer et al., 2003). Nisbett et al. (2012) faz uma revisão das inúmeras influências que variações ambientais exercem sobre as habilidades cognitivas.

[121] Diversos dos parágrafos subsequentes são amplamente baseados em um trabalho realizado em conjunto com Carl Shulman (Shulman e Bostrom, 2014).

[122] Com informações imprecisas sobre os efeitos genéticos aditivos na habilidade cognitiva, o tamanho desses efeitos seria reduzido. Entretanto, até mesmo um pouco de conhecimento poderia gerar resultados expressivos, visto que os ganhos com a seleção não crescem de forma linear com a proporção da variância que podemos explicar. Ao contrário, a efetividade da seleção depende do desvio padrão do QI médio previsto, que aumenta proporcionalmente à *raiz quadrada* da variância. Por exemplo, se fôssemos capazes de explicar 12,5% da variância, isso poderia gerar efeitos tão grandes quanto a metade dos apresentados na tabela 1, que levam em consideração um valor de 50%. Comparativamente, um estudo recente (Rietveld et al., 2013) afirma já ter identificado 2,5% da variância.

[123] Para efeito de comparação, a prática padrão atual envolve a criação de menos de dez embriões.

[124] Células-tronco adultas e embrionárias podem ser estimuladas a se desenvolver em um espermatozoide ou em um ovócito, que, por sua vez, podem ser fundidos para gerar um embrião (Nagy et al., 2008; Nagy e Chang, 2007). Os percursores dos óvulos também podem formar blastocistos partenogenéticos, embriões não fertilizados e não viáveis capazes de produzir uma linha de células-tronco embrionárias para o processo em questão (Mai et al., 2007).

[125] Essa é a opinião de Katsuhiko Hayashi, relatada em Cyranoski (2013). O Hinxton Group, um consórcio internacional de cientistas que discutem a respeito de questões éticas e desafios nas pesquisas com células-tronco, estimou, em 2008, que células-tronco humanas, derivadas de gametas, estariam disponíveis em dez anos (Hinxton Group, 2008), e os desenvolvimentos nas pesquisas até o momento são bastante condizentes com essa previsão.

[126] Sparrow (2013); Miller (2012); *The Uncertain Future* (2012).

[127] Sparrow (2013).

[128] As preocupações de cunho mais laico provavelmente teriam como foco antecipar os impactos dessa tecnologia em questões como a desigualdade social, a segurança médica do procedimento, o medo de que aconteça uma “corrida dos ratos” em busca do aperfeiçoamento, os direitos e responsabilidades dos pais em relação às crianças, as sombras, ainda presentes, das teorias eugenistas do século XX, o conceito de dignidade humana e os limites apropriados para a intervenção do Estado nos direitos reprodutivos de seus cidadãos. (Para uma discussão a respeito da ética nas melhorias cognitivas, ver Bostrom e Ord [2006], Bostrom e Roache [2011] e Sandberg e Savulescu [2011].) Algumas tradições religiosas podem trazer preocupações adicionais,

entre elas aquelas centradas no status moral dos embriões ou nos limites apropriados para a ação humana no plano da criação.

[129] Para evitar os efeitos negativos da endogamia, o processo iterativo de seleção de embriões iria requerer um imenso suprimento inicial de doadores ou um gasto substancial de poder seletivo a fim de reduzir alelos recessivos nocivos. Ambas as alternativas teriam uma tendência de fazer com que os filhos fossem menos semelhantes geneticamente a seus pais (e mais semelhantes uns com os outros).

[130] O quão difícil será o obstáculo imposto pela epigenética ainda é uma questão desconhecida (Chason et al., 2011; Iliadou et al., 2011).

[131] Embora as habilidades cognitivas sejam geralmente hereditárias, talvez existam alguns poucos alelos incomuns ou polimorfismos que venham a ter, individualmente, um amplo efeito positivo na inteligência (Davis et al., 2010; Davies et al., 2011; Rietveld et al., 2013). Conforme os métodos de sequenciamento sejam aprimorados, será cada vez mais factível o mapeamento de alelos de baixa frequência, o que possibilitará, por sua, vez a compreensão de sua correlação com o desenvolvimento da cognição e do comportamento. Algumas evidências teóricas sugerem que alguns alelos responsáveis por distúrbios genéticos em homozigotos podem oferecer consideráveis vantagens cognitivas nos portadores heterozigóticos, levando a um prognóstico de que os heterozigotos presentes nas doenças de Gaucher, Tay-Sachs e Niemann-Pick teriam um QI cinco pontos maior do que os dos grupos de controle (Cochran et al., 2006). O tempo dirá se tal prognóstico se sustenta.

[132] Um artigo (Nachman e Crowell, 2000) estima que ocorram cerca de 175 mutações por genoma a cada geração. Outro artigo (Lynch, 2010) estima, através de uma metodologia diferente, que um recém-nascido normal tenha entre cinquenta e cem mutações, e Kong et al. (2012) sugere um número de cerca de 77 novas mutações por geração. A maioria dessas mutações não afeta o funcionamento do organismo, ou o faz em grau praticamente imperceptível, mas o efeito combinado de diversas mutações prejudiciais poderia significar uma perda considerável de aptidão. Ver também Crow (2000).

[133] Crow (2000); Lynch (2010).

[134] É necessário fazer algumas ressalvas importantes em relação a essa ideia. Possivelmente, o genoma modal precisaria de alguns ajustes para que alguns problemas fossem evitados. Por exemplo, partes do genoma poderiam ser adaptadas para interagir com outras partes, partindo-se do princípio de que todas as partes funcionem com um certo grau de eficiência. O aumento de eficiência dessas partes poderia, então, nos levar a ultrapassar algumas etapas metabólicas.

[135] As intervenções poderiam causar um impacto mais imediato, por exemplo, modificando a expectativa das pessoas em relação ao futuro.

[136] Louis Harris & Associates (1969); Mason (2003).

[137] Kalfoglou et al. (2004).

[138] Os dados são obviamente limitados, mas em testes de habilidade infantil, indivíduos com resultados excepcionais (1 em 10 mil) mostraram, em estudos longitudinais, possuir uma capacidade muito maior de se tornar professores universitários, desenvolver patentes científicas e ser bem-sucedidos nos negócios do que aqueles indivíduos com resultados um pouco menos excepcionais nos testes (Kell et al., 2013). Roe (1953) fez um estudo com 64 cientistas eminentes e descobriu neles uma habilidade cognitiva mediana três a quatro vezes maior, considerado o desvio padrão, do que a tipicamente encontrada em cientistas no geral. (A habilidade cognitiva também tem correlação com os rendimentos monetários ao longo da vida e com um melhor desempenho em outras áreas, como expectativa de vida, taxa de divórcio e chance de abandonar os estudos [Deary, 2012].) Um deslocamento positivo na distribuição de habilidades cognitivas teria efeitos muito desproporcionais nas extremidades, especialmente porque aumentaria o número de pessoas superdotadas e reduziria o número de pessoas com retardos mentais e distúrbios de aprendizagem. Ver também Bostrom e Ord (2006) e Sandberg e Savulescu (2011).

[139] Como exemplo, ver Warwick (2002). Stephen Hawking chegou a sugerir que esse seria um passo necessário para que nos fosse possível controlar os avanços em inteligência de máquina: “Nós devemos, o mais rapidamente possível, desenvolver tecnologias que tornem possível uma conexão direta entre o cérebro e o computador, de forma que os cérebros artificiais contribuam com a inteligência humana em vez de se opor a ela” (citado em Walsh [2001]). Ray Kurzweil concorda: “Com relação à recomendação de Hawking […] em particular no que tange a conexão direta entre cérebro e computador, concordo que isso é ao mesmo tempo razoável, desejável e inevitável. [*sic*] Por anos essa tem sido a minha recomendação” (Kurzweil, 2001).

[140] Ver Lebedev e Nicolelis (2006); Birbaumer et al. (2008); Mak e Wolpaw (2009); e Nicolelis e Lebedev (2009). Uma perspectiva mais pessoal a respeito do problema de aperfeiçoamento de seres humanos através de implantes pode ser encontrada em Chorost (2005, cap. 11).

[141] Smeding et al. (2006).

[142] Degnan et al. (2002).

[143] Dagnelie (2012); Shannon (2012).

[144] Perlmutter e Mink (2006); Lyons (2011).

[145] Koch et al. (2006).

[146] Schalk (2008). Para uma visão geral do estado da arte, ver Berger et al. (2008). Para uma argumentação que afirma que isso ajudaria a levar a um aperfeiçoamento da inteligência, ver Warwick (2002).

[147] Alguns exemplos: Bartels et al. (2008); Simeral et al. (2011); Krusienski e Shih (2011); e Pasqualotto et al. (2012).

[148] Como exemplo, ver Hinke et al. (1993).

[149] Existem algumas exceções parciais, especialmente nos primeiros estágios dos processos sensoriais. Por exemplo, o córtex visual primário utiliza um mapeamento retinotópico, o que significa dizer que grupos neurais adjacentes recebem dados de áreas adjacentes da retina (embora as colunas de dominância ocular compliquem um pouco esse mapeamento).

[150] Berger et al. (2012); Hampson et al. (2012)

[151] Alguns implantes cerebrais requerem duas formas de aprendizagem: o dispositivo precisa aprender a interpretar as representações neurais do organismo e este, por sua vez, precisa aprender a usar o sistema por meio da criação de padrões apropriados de ativação neural (Carmena et al., 2003).

[152] Tem-se sugerido que deveríamos ver as entidades corporativas (corporações, sindicatos, governos, igrejas etc.) como agentes com inteligência artificial, entidades com sensores e atuadores, capazes de representar conhecimento, fazer inferências e agir (ver exemplos em Kuipers [2012] e Huebner [2008] para uma discussão a respeito da existência, ou não, de representações coletivas). Elas são claramente poderosas e ecologicamente bem-sucedidas, apesar de suas capacidades e estados internos encontrarem diferenças em relação àqueles dos seres humanos.

[153] Hanson (1995, 2000); Berg e Rietz (2003).

[154] No ambiente de trabalho, por exemplo, os empregadores poderiam utilizar detectores de mentira para coibir roubos e desvios por parte dos funcionários, bastando apenas perguntar, ao fim de cada jornada de trabalho, se o funcionário roubou alguma coisa ou se trabalhou o máximo que podia. Do mesmo modo, líderes políticos e empresariais poderiam ser questionados a fim de verificar se eles estão representando genuinamente os interesses de seus acionistas ou eleitores. Ditadores poderiam utilizar os detectores de mentira na procura de generais insurgentes dentro do regime ou elementos subversivos na população.

[155] É possível imaginar o surgimento de técnicas de neuroimageamento capazes de detectar padrões neurais de cognição motivada. Sem um mecanismo de detecção de autoengano, os detectores de mentiras favoreceriam indivíduos que acreditam em seus próprios ideais, mesmo que equivocados. Testes melhores para o autoengano também poderiam ser usados no treinamento da racionalidade e no estudo da efetividade de intervenções que objetivam a redução de preconceitos.

[156] Bell e Gemmel (2009). Um dos primeiros exemplos a esse respeito pode ser encontrado no trabalho de Deb Roy, do Instituto de Tecnologia de Massachusetts (MIT), que gravou cada momento dos primeiros três de anos de vida de seu filho. A análise desses dados audiovisuais fornece informações para pesquisas acerca do desenvolvimento da linguagem; ver Roy (2012).

[157] O aumento puro e simples da população mundial contribuirá muito pouco. Cenários que envolvessem inteligência de máquina poderiam levar a um aumento da população mundial (inclusive as mentes digitais) de muitas ordens de magnitude e em um curto espaço de tempo. Mas esse caminho até a superinteligência envolve IA ou emulação completa do cérebro, então, iremos desconsiderá-lo aqui.

[158] Vinge (1993).

[159] Vernor Vinge usou o termo “superinteligência fraca” para se referir a tais mentes humanas aceleradas (Vinge, 1993).

[160] Por exemplo, se um sistema superveloz pudesse fazer tudo o que qualquer humano pode fazer exceto dançar mazurca, ainda assim deveria ser considerado uma superinteligência rápida. Nosso interesse está nas capacidades cognitivas centrais que possuam significado econômico ou estratégico.

[161] Uma aceleração pelo menos 1 milhão de vezes maior do que os cérebros humanos é fisicamente possível, como se pode ver considerando a diferença de velocidade e energia requeridas por processos cerebrais relevantes em comparação com um processamento de informação mais eficiente. A velocidade da luz é mais do que 1 milhão de vezes maior que a velocidade de transmissão neural, os disparos sinápticos dissipam 1 milhão de vezes mais calor do que é termodinamicamente necessário e as frequências atuais dos transistores são 1

milhão de vezes mais rápidas que as frequências de disparo dos neurônios (Yudkowsky [2008a]; ver também Drexler [1992]). Os limites máximos da superinteligência rápida são definidos pelos atrasos de comunicação à velocidade da luz, pelos limites quânticos na velocidade de transição de estados e pelo volume necessário para conter a mente (Lloyd, 2000). O "laptop de ponta" descrito por Lloyd (2000) rodaria uma emulação cerebral a $1{,}4\times10^{21}$ FLOPS com aceleração de $3{,}8\times10^{29}$ × (assumindo que a emulação pudesse ser suficientemente paralelizada). O projeto de Lloyd, entretanto, não tem a intenção de ser tecnologicamente plausível; seu objetivo é apenas ilustrar as limitações na computação que são facilmente derivadas das leis básicas da física.

[162] Com as emulações, há também a questão de por quanto tempo uma mente similar à humana poderia continuar a trabalhar em algo antes de enlouquecer ou cair numa rotina. Mesmo com uma variedade de tarefas e férias regulares, não é óbvio que uma mente similar à humana possa viver por milhares de anos subjetivos sem desenvolver problemas psicológicos. Além do mais, se a capacidade total de memória é limitada — uma consequência de haver uma população limitada de neurônios —, então um aprendizado cumulativo não pode continuar indefinidamente: a partir de algum ponto, a mente precisa começar a esquecer uma coisa para cada coisa nova que ela aprender. (A inteligência artificial pode ser projetada para amenizar esses eventuais problemas.)

[163] Consequentemente, nanomecanismos se movendo na modesta velocidade de 1 m/s têm escalas típicas de tempo de nanossegundos. Ver a seção 2.3.2 de Drexler (1992). Robin Hanson se refere a robôs de 7-mm "Tinkerbell" se deslocando 260 vezes mais rápido que velocidade normal (Hanson, 1994).

[164] Hanson (2012).

[165] "Inteligência coletiva" não se refere à paralelização em baixo nível do hardware dos computadores, mas à paralelização no nível de inteligência de agentes autônomos, tais como os seres humanos. Implementar uma única emulação em uma máquina maciçamente paralela talvez resulte em uma superinteligência rápida se o computador paralelo for suficientemente rápido, mas não produziria uma inteligência coletiva.

[166] Melhorias na velocidade ou na qualidade dos componentes individuais poderiam também afetar indiretamente o desempenho da inteligência coletiva, mas aqui consideramos principalmente tais melhorias sob outras duas formas de superinteligência em nossa classificação.

[167] Alguns argumentos afirmam que uma densidade mais alta da população teria provocado a revolução do Paleolítico Superior e que, a partir de um determinado limiar, o acúmulo da complexidade cultural se tornou muito mais facilitado (Powell et al., 2009).

[168] E em relação à internet? Parece que ela ainda não resultou em um impulso gigante. Talvez isso venha a acontecer, por fim. Foram necessários séculos ou milênios para que os outros exemplos listados aqui revelassem todo o seu potencial.

[169] A intenção não é, obviamente, que este seja um exercício intelectual realístico. Um planeta grande o suficiente para sustentar 7 quatrilhões de organismos humanos com tecnologia atual implodiria, a menos que fosse feito de um material muito leve ou fosse oco e sustentado por pressão ou outros meios artificiais. (Uma esfera de Dyson ou um grande planetoide artificial composto de diversos níveis esféricos concêntricos poderiam ser soluções melhores.) A história teria se desenrolado de modo diferente numa vasta superfície como essa. Ignore tudo isso.

[170] Focamos, aqui, nas propriedades funcionais de um intelecto unificado, e não na possibilidade de tal intelecto possuir *qualia* ou se ele seria uma mente no sentido de ter uma experiência subjetiva consciente. (É possível conjecturar, entretanto, quais tipos de experiências conscientes podem surgir de intelectos que sejam mais ou menos integrados do que os cérebros humanos. Em algumas perspectivas sobre a consciência, tais como a Teoria do Espaço de Trabalho Global, pode-se, aparentemente, esperar que cérebros mais integrados tenham mais capacidade de consciência. Cf. Baars [1997]; Shanahan [2010]; e Schwitzgebel [2013].)

[171] Mesmo pequenos grupos de humanos que permaneçam isolados por algum tempo ainda podem se beneficiar dos resultados de uma inteligência coletiva maior. Por exemplo, a linguagem usada por eles pode ter sido desenvolvida por uma comunidade linguística mais ampla e suas ferramentas podem ter sido inventadas num contexto com uma população maior antes que o grupo menor se isolasse. Mas ainda que tenha sempre permanecido isolado, um grupo pequeno pode fazer parte de uma inteligência coletiva maior por mais difícil que isso possa parecer — ou seja, a inteligência coletiva consiste não apenas do presente, mas também de todas as gerações ancestrais, constituindo um agregado que pode funcionar como um sistema de processamento e transmissão de informações.

[172] De acordo com a tese de Church-Turing, todas as funções computáveis são computáveis por uma máquina de Turing. Como qualquer uma das três formas de superinteligência poderia simular uma máquina de Turing (se tivesse acesso a uma memória ilimitada e pudesse operar por tempo indefinido), elas são, segundo esse critério formal, computacionalmente equivalentes. Certamente, um ser humano comum (provido de uma quantidade ilimitada de papel para rascunho e tempo ilimitado à disposição) poderia também implementar uma

máquina de Turing e, portanto, é também equivalente pelo mesmo critério. O que interessa para os nossos propósitos, entretanto, é o que esses sistemas diferentes podem conseguir *na prática* com uma memória finita e em um tempo razoável. E as variações de eficiência são tão grandes que é possível fazer algumas distinções prontamente. Por exemplo, um típico indivíduo com um QI de 85 poderia ser ensinado a implementar uma máquina de Turing. (Talvez seja até possível treinar um chimpanzé particularmente dócil e inteligente para fazer isso.) Mesmo assim, para todas as intenções e propósitos práticos, tal indivíduo seria presumidamente incapaz de, digamos, desenvolver de forma independente a teoria geral da relatividade ou de ganhar uma medalha Fields.

[173] As tradições de contar histórias oralmente podem produzir ótimos trabalhos (tais como as epopeias de Homero), mas talvez alguns dos autores contribuintes possuíssem talentos incomuns nesses casos.

[174] A menos que ela seja composta por intelectos que tenham superinteligência rápida ou de qualidade.

[175] A nossa incapacidade de especificar todos esses problemas pode, em parte, ser decorrente da falta de tentativa: não há muito sentido em passar o tempo detalhando tarefas intelectuais que nenhum indivíduo ou organização atualmente existentes seja capaz de realizar. Mas também é possível que mesmo o ato de conceitualizar algumas dessas tarefas seja, por si só, uma dessas atividades cuja realização demande cérebros que não possuímos.

[176] Cf. Boswell (1917); ver também Walker (2002).

[177] Essa frequência ocorre principalmente em descargas curtas de um subconjunto de neurônios — a maioria tem um ritmo de disparos mais lento (Gray e McCormick, 1996; Steriade et al., 1998). Há alguns neurônios (“neurônios tagarelas”, também conhecidos como células “ritmicamente explosivas”) que podem alcançar frequências de disparo de até 750 Hz, mas eles parecem ser extremamente discrepantes.

[178] Feldman e Ballard (1982).

[179] A velocidade de condução depende do diâmetro do axônio (axônios mais grossos são mais rápidos) e se este é mielinizado ou não. Dentro do sistema nervoso central, atrasos de transmissão podem variar de menos de um milissegundo até 100 ms (Kandel et al., 2000). A transmissão via fibra ótica é por volta de 68% *c* (por causa do índice de refração do material). Cabos elétricos têm aproximadamente a mesma velocidade, 59-77% *c*.

[180] Supondo uma velocidade de sinal de 70% *c*. Se a velocidade for 100% *c*, a estimativa é aumentada para $1{,}8\times10^{18}$ m³.

[181] A quantidade de neurônios presentes no cérebro de um homem adulto foi estimada em 86,1 ± 8,1 bilhões, um número alcançado por meio da dissolução de cérebros e separação dos núcleos das células, contando-se aqueles tingidos com um marcador específico para neurônios. No passado, estimativas próximas a 75-125 bilhões eram comuns. Estas eram tipicamente baseadas na contagem manual das densidades de células em áreas representativas pequenas (Azevedo et al., 2009).

[182] Whitehead (2003).

[183] Sistemas de processamento de informação podem certamente usar processos em escala molecular para computação e armazenamento de dados e alcançar pelo menos um tamanho planetário em extensão. Os limites físicos definitivos para a computação estabelecidos pela mecânica quântica, relatividade geral e termodinâmica, entretanto, vão além do nível desse “cérebro de Júpiter” (Sandberg, 1999; Lloyd, 2000).

[184] Stansberry e Kudritzki (2012). A eletricidade usada em *data centers* mundialmente equivale a 1,1-1,5% da eletricidade total utilizada (Koomey, 2011). Ver também Muehlhauser e Salamon (2012).

[185] Essa é uma simplificação excessiva. O número de elementos que a memória de trabalho pode manter é dependente tanto do tipo de informação que eles contêm quanto do tipo da tarefa que está sendo executada. Entretanto, a memória de trabalho é claramente limitada a um número pequeno de elementos. Ver Miller (1956) e Cowan (2001).

[186] Um exemplo talvez seja que a dificuldade de aprender conceitos booleanos (categorias definidas por regras lógicas) é proporcional ao tamanho da fórmula proposicional logicamente equivalente mais curta. Normalmente, até fórmulas com apenas 3-4 literais de comprimento são muito difíceis de aprender. Ver Feldman (2000).

[187] Ver Landauer (1986). Esse estudo é baseado em estimativas experimentais da taxa de aprendizado e esquecimento nos humanos. Se levarmos em conta o aprendizado implícito, talvez a estimativa aumente um pouco. Se considerarmos uma capacidade de armazenamento de ~1 bit por sinapse, é possível obter um *limite superior* de capacidade da memória humana de cerca de 1015 bits. Para uma visão geral de estimativas diferentes, ver o “Apêndice A” de Sandberg e Bostrom (2008).

[188] A interferência de canal pode desencadear potenciais de ação, e o ruído sináptico produz uma variabilidade significativa na força dos sinais transmitidos. Os sistemas nervosos parecem ter evoluído para fazer diversas compensações entre tolerância ao ruído e custos (massa, tamanho, atrasos); ver Faisal et al. (2008). Por exemplo, axônios não podem ser menores do que 0,1 mm para que nenhuma abertura aleatória dos canais de íon crie potenciais de ação espontâneos (Faisal et al., 2005).

[189] Trachtenberg et al. (2002).

[190] Em termos de memória e poder computacional, embora não em termos de eficiência energética. O computador mais rápido do mundo na época que este livro foi escrito era o computador chinês Tianhe-2, que desbancou o Titan da Cray Inc. em junho de 2013 com um desempenho de 33,86 petaFLOPS. Essa máquina consome 17,6 MW de energia, quase seis ordens de magnitude acima dos ~20 W do cérebro.

[191] Note que esse levantamento sobre as vantagens da máquina é *disjuntivo*: nosso argumento continua válido mesmo que algum dos itens listados se provem ilusórios, desde que haja pelo menos um item que possa proporcionar uma vantagem suficientemente grande.

[192] O sistema pode não alcançar uma dessas referências em um momento bem definido. Em vez disso, é possível que haja um intervalo durante o qual o sistema gradualmente se torne capaz de superar uma equipe externa de pesquisa em um número crescente de tarefas de desenvolvimento de melhorias do sistema.

[193] Na última metade do século passado, pelo menos um cenário no qual a ordem mundial existente acabaria em questão de minutos ou horas foi amplamente identificado: uma guerra termonuclear global.

[194] Isso seria consistente com a observação de que o efeito de Flynn — o aumento secular na medida de QI na maioria das populações em uma taxa de aproximadamente 3 pontos por década nos últimos sessenta anos, aproximadamente — parece ter cessado ou até mesmo sofrido retrocessos recentemente em alguns países altamente desenvolvidos como o Reino Unido, a Dinamarca e a Noruega (Teasdale e Owen, 2008; Sundet et al., 2004). A causa do efeito de Flynn no passado — e se e em que medida ele representa qualquer ganho genuíno em inteligência geral ou é meramente um aumento da habilidade na resolução de testes de QI — tem sido tema de amplos debates e ainda não é conhecida. Mesmo que o efeito de Flynn (pelo menos parcialmente) reflita ganhos cognitivos reais, e ainda que o efeito esteja atualmente diminuindo ou até regredindo, isso não prova que tenhamos chegado num ponto em que há uma diminuição dos retornos em quaisquer que sejam as causas responsáveis pela observação do efeito de Flynn no passado. O declínio ou regressão poderiam, ao contrário, ser devidos a algum fator prejudicial independente que supostamente teria produzido um declínio ainda maior caso não houvesse o efeito de Flynn.

[195] Bostrom e Roache (2011).

[196] A terapia genética somática poderia eliminar o atraso de maturação, mas é tecnicamente muito mais desafiadora do que intervenções nas linhas germinais e possui um potencial menor.

[197] A média do crescimento anual da produtividade global entre o período de 1960-2000 foi de 4,3% (Isaksson, 2007). Apenas uma parte desse crescimento na produtividade é devida a ganhos em eficiência organizacional. Algumas redes ou processos organizacionais *em particular* estão, com certeza, melhorando muito mais rapidamente.

[198] A evolução de cérebros biológicos foi sujeita a várias limitações e compensações que podem ser drasticamente atenuadas quando a mente passa para um meio digital. Por exemplo, o tamanho do cérebro é limitado pelo tamanho da cabeça, e se a cabeça for grande demais, é impossível fazê-la passar pelo canal do parto. Um cérebro grande também utiliza muitos recursos metabólicos e é um peso morto que dificulta o movimento. A conectividade entre certas regiões do cérebro pode ser limitada por restrições espaciais — o volume de massa branca é significantemente maior do que o volume de massa cinzenta que ela conecta. A dissipação de calor é limitada pelo fluxo de sangue e pode estar próxima ao limite superior para um funcionamento aceitável. Além disso, os neurônios biológicos têm ruído, são lentos e precisam de proteção constante, manutenção e reabastecimento por células gliais e vasos sanguíneos (o que contribui para uma aglomeração na região intracraniana). Ver Bostrom e Sandberg (2009b).

[199] Yudkowsky (2008a, p. 326). Para uma discussão mais recente, ver Yudkowsky (2013).

[200] Lin et al. (2012).

[201] Um certo aumento da inteligência coletiva pode ser alcançado simplesmente com a ampliação do número de seus intelectos constituintes. Isso deveria pelo menos permitir um desempenho geral melhor nas tarefas que possam ser facilmente paralelizadas. Todavia, para tirar proveito máximo de tal explosão populacional, também seria necessário obter algum nível de coordenação (maior do que o mínimo) entre os constituintes.

[202] De qualquer maneira, a distinção entre a inteligência rápida e a inteligência de qualidade não é tão clara no caso dos sistemas de IA não neuromórficos.

[203] Rajab et al. (2006, p. 41-52).

[204] Tem sido sugerido que o uso de circuitos integrados configuráveis (FPGAS), em vez de processadores de propósito geral, poderia aumentar a velocidade computacional em simulações de redes neurais em até duas ordens de magnitude (Markram, 2006). Um estudo de modelos climáticos de alta resolução com necessidades computacionais da ordem de petaFLOPS observou uma redução de custo de 24 a 32 vezes e uma redução de duas ordens de magnitude na necessidade energética com a utilização de versões customizadas de processadores embarcados (Wehner et al., 2008).

[205] Nordhaus (2007). Diversos artigos fornecem um panorama dos diferentes significados da lei de Moore; ver, por exemplo, Tuomi (2002) e Mack (2011).

[206] Se o desenvolvimento for lento o suficiente, o projeto poderá se beneficiar dos progressos que ocorram mundo afora nesse meio-tempo, como os avanços na ciência da computação alcançados por pesquisadores em universidades e melhorias no hardware desenvolvido pela indústria de semicondutores.

[207] Um excedente algorítmico talvez seja menos provável, mas uma exceção aconteceria se um hardware diferente, tal como um computador quântico, se tornasse disponível para rodar algoritmos que não eram viáveis previamente. Pode-se também argumentar que redes neurais e *deep learning* são casos de excedente algorítmico: computacionalmente muito caros para funcionar bem no momento em que foram inventados, eles foram engavetados por um bom tempo, sendo mais tarde utilizados quando unidades de processamento gráfico (GPUS) suficientemente rápidas puderam baratear seu custo computacional. Atualmente, eles vencem competições.

[208] E inclusive se o progresso em direção à linha de referência humana fosse lento.

[209] Algumas tentativas experimentais têm sido feitas para desenvolver a ideia de uma explosão de inteligência dentro dos moldes da teoria de crescimento econômico; ver, por exemplo, Hanson (1998b); Jones (2009); Salamon (2009). Esses estudos têm indicado o potencial de crescimento extremamente rápido possibilitado pela chegada das mentes digitais, mas já que a teoria do crescimento endógeno é relativamente mal desenvolvida mesmo para aplicações históricas e contemporâneas, qualquer aplicação num contexto futuro potencialmente descontínuo é melhor vista nesse estágio como uma fonte de conceitos e considerações potencialmente úteis do que como um exercício que provavelmente traria previsões confiáveis. Para uma visão sobre as tentativas de modelar matematicamente uma singularidade tecnológica, ver Sandberg (2010).

[210] É também possível que não haja nenhuma partida. Mas já que, conforme discutido anteriormente, a superinteligência parece tecnicamente plausível, a ausência de uma partida se deveria provavelmente à intervenção de algum obstáculo, tal como uma grande catástrofe. Se a superinteligência forte surgisse não na forma de inteligência artificial ou de uma emulação completa do cérebro, mas através de um dos outros caminhos que consideramos acima, então uma partida lenta seria mais provável.

[211] O_{mundo} representa a parte do poder de otimização mundial aplicada para melhorar o sistema em questão. Para um projeto que funcione em isolamento completo, sem receber apoio significativo constante do mundo externo, teríamos $O_{mundo} \approx 0$, mesmo considerando que o projeto certamente teria começado com recursos (computadores, conceitos científicos, pessoal com formação adequada etc.) derivados da economia mundial e muitos séculos de desenvolvimento.

[212] Aqui, a habilidade cognitiva mais relevante de uma IA embrionária é a capacidade de criar projetos inteligentes para melhorar a si mesma, ou seja, sua capacidade de amplificar a própria inteligência. (Se a IA embrionária for boa em melhorar outro sistema, que, por sua vez, seja bom em melhorar a IA embrionária, então nós poderíamos vê-los como subsistemas de um sistema maior e concentraríamos nossas análises no todo.)

[213] Supondo que a resistência não seja considerada tão alta a ponto de desestimular investimentos por completo ou desviá-los para algum projeto alternativo.

[214] Um exemplo similar é abordado em Yudkowsky (2008b).

[215] Como os investimentos aumentaram (por exemplo, quantidades investidas em construções de novas fábricas e o número de pessoas trabalhando na indústria de semicondutores), a lei de Moore por si só não gerou um crescimento tão rápido descontando-se o efeito do crescimento nos investimentos. Todavia, em combinação com avanços de software, uma duplicação de desempenho por unidade de investimento a cada dezoito meses pode ser mais plausível historicamente.

[216] Uma mente de software pode rodar em uma única máquina em vez de usar toda a rede mundial de computadores; mas isso não é o que nós queremos dizer por “concentração”. O que nos interessa aqui é a extensão pela qual o poder, especificamente o poder derivado de habilidades tecnológicas, estará concentrado nos estágios avançados ou imediatamente após a revolução da inteligência de máquina.

[217] A difusão tecnológica de produtos de consumo, por exemplo, tende a ser mais lenta em países em desenvolvimento (Talukdar et al., 2002). Ver também Keller (2004) e The World Bank (2008).

[218] A literatura econômica que lida com a teoria da firma é relevante como ponto de comparação para a atual discussão. O texto clássico é Coase (1937). Ver também, por exemplo, Canbäck et al. (2006); Milgrom e Roberts (1990); Hart (2008); Simester e Knez (2002).

[219] Por outro lado, poderia ser especialmente fácil roubar uma IA embrionária, já que ela seria um software que poderia ser transmitido eletronicamente ou copiado num dispositivo portátil de memória.

[220] Barber (1991) sugere que a cultura Yangshao (5000-3000 a.C.) pode ter feito uso da seda. Sun et al. (2012) estimam, com base em estudos genéticos, que a domesticação do bicho-da-seda tenha ocorrido cerca de 4.100 anos atrás.

[221] Cook (1984, p. 144). Pode ser que essa história seja boa demais para resistir a uma investigação histórica, assim como a história de Procópio (*Wars*, VIII, XVII, p. 1-7) sobre como os bichos-da-seda foram supostamente levados até Bizâncio por monges andarilhos que os escondiam em suas varas de bambus ocas (Hunt, 2011).

[222] Wood (2007); Temple (1986).

[223] As culturas pré-colombianas conheciam a roda, mas a usavam apenas em brinquedos (provavelmente em virtude da falta de bons animais de carga).

[224] Koubi (1999); Lerner (1997); Koubi e Lalman (2007); Zeira (2011); Judd et al. (2012).

[225] Estimado a partir de uma série de fontes. O intervalo de tempo, de alguma forma, é geralmente arbitrário, dependendo exatamente de que maneira as capacidades “equivalentes” são definidas. O radar foi utilizado por pelo menos dois países nos primeiros anos em que foi introduzido, mas o número exato em meses é difícil de calcular.

[226] Ellis (1999).

[227] Em 1953, o RDS-6 foi o primeiro teste de uma bomba com reações de fusão, mas, em 1955, o RDS-37 foi a primeira bomba de fusão “verdadeira”, onde grande parte da força veio da reação de fusão.

[228] Não confirmado.

[229] Testes em 1989, projeto cancelado em 1994.

[230] Sistema instalado, capaz de alcançar mais de 5.000 km.

[231] Mísseis Polaris comprados dos Estados Unidos.

[232] Atualmente há um trabalho em andamento no míssil Taimur, possivelmente baseado nos mísseis chineses.

[233] O foguete RSA-3, testado em 1989-90, foi projetado para lançamentos de satélites e/ou como um ICBM.

[234] MIRV, da sigla em inglês para Mísseis de Reentrada Múltipla Independentemente Direcionados, tecnologia que permite que um único míssil balístico carregue múltiplas ogivas que podem ser programadas para atingir alvos diferentes.

[235] O sistema Agni V ainda não se encontra em funcionamento.

[236] Se a diferença de tempo entre os projetos for modelada como uma distribuição normal, então a distância provável entre o projeto líder e seu seguidor mais próximo também dependerá de quantos projetos existam. Se houver um grande número de projetos, então a distância entre os dois primeiros é provavelmente pequena, mesmo se a variância da distribuição for moderadamente alta (embora a distância esperada entre o projeto líder e o segundo projeto caia muito lentamente quando o número de competidores aumenta, se os tempos de finalizações forem normalmente distribuídos). Entretanto, é improvável que haja um grande número de projetos e que cada um seja financiado suficientemente bem para o surgimento de concorrentes sérios. (Poderá haver um número maior de projetos se houver um grande número de abordagens básicas diferentes que possam ser seguidas, mas nesse caso é provável que muitas dessas abordagens não cheguem a lugar nenhum.) Como sugerido, empiricamente parece não haver, em geral, mais do que alguns poucos competidores sérios buscando qualquer objetivo tecnológico específico. A situação é um pouco diferente em mercados de consumo onde há muitos nichos para diferentes produtos e onde existem poucas barreiras para a entrada. Há muitos projetos individuais para a produção de camisetas, mas apenas algumas empresas no mundo estão desenvolvendo a nova geração de placas de vídeo. (Duas empresas, AMD e NVIDIA, são quase um duopólio no momento, embora a Intel também esteja competindo nesse mercado num segmento de menor performance.)

[237] Bostrom (2006c). Pode-se imaginar um *singleton* cuja existência seja invisível (por exemplo, uma superinteligência com tecnologia e discernimento tão avançados que poderia sutilmente controlar eventos no mundo sem que qualquer humano percebesse essa intervenção); ou um *singleton* que voluntariamente impusesse limitações muito rigorosas em seu próprio exercício de poder (por exemplo, limitar meticulosamente a si mesmo para assegurar que certas regras definidas por tratados internacionais — ou princípios libertários — sejam respeitadas). A probabilidade de que um tipo particular de *singleton* surja é, com certeza, uma questão

empírica; mas *conceitualmente*, pelo menos, é possível haver um *singleton* bom, um *singleton* ruim, um *singleton* extravagantemente diferente, um *singleton* suavemente monolítico, um *singleton* castradoramente opressivo ou um *singleton* que se pareça mais com uma lei extra da natureza do que com um déspota autoritário.

[238] Jones (1985, p. 344).

[239] O fato de que o Projeto Manhattan tenha sido desenvolvido durante um período de guerra poderia ser significativo. Muitos dos cientistas que participaram afirmaram estar primeiramente motivados pela situação do momento de guerra e pelo medo de que a Alemanha nazista pudesse desenvolver armas nucleares antes dos aliados. Pode ser difícil para muitos governos mobilizar um esforço igualmente secreto e intenso em tempos de paz. O programa Apollo, outro megaprojeto icônico da ciência e engenharia, recebeu um forte ímpeto da rivalidade da Guerra Fria.

[240] Embora, mesmo que *estivessem* procurando com afinco, não seja claro se as agências deixariam que isso transparecesse (publicamente).

[241] Técnicas criptográficas possibilitariam que a equipe de colaboradores estivesse fisicamente dispersa. O único elo fraco na corrente de comunicação poderia ser o estágio de entrada de dados, já que o ato físico de teclar poderia ser observado. Mas se a vigilância interna se tornasse comum (por meio de aparelhos de gravação microscópicos), indivíduos muito interessados em proteger sua privacidade poderiam desenvolver contramedidas (por exemplo, armários especiais que fornecessem blindagem contra dispositivos potencialmente espiões). Considerando-se que o espaço físico pode se tornar transparente numa futura era de vigilância, o espaço cibernético também poderia ser mais protegido por meio da ampla adoção de protocolos criptográficos mais rígidos.

[242] Um estado totalitário poderia recorrer até mesmo a medidas mais coercitivas. Cientistas que estivessem atuando nas aéreas mais relevantes poderiam ser capturados e aprisionados em campos de trabalho semelhantes às “vilas acadêmicas” da Rússia stalinista.

[243] Se o nível de preocupação pública for relativamente baixo, alguns pesquisadores talvez possam receber bem um pouco de alarmismo, pois isso chama a atenção para o seu trabalho e faz com que seu campo de pesquisa pareça importante e excitante. Quando o nível de preocupação se torna maior, as comunidades relevantes de pesquisa podem mudar de opinião conforme comecem a se preocupar com cortes de financiamento, regulamentações e resistência da opinião pública. Pesquisadores em disciplinas afins — tais como as áreas de ciência da computação e robótica, que não sejam muito relevantes para a inteligência artificial geral — poderiam se ressentir do desvio de financiamento e atenção das suas próprias áreas de pesquisas. Esses profissionais também podem observar com razão que *seus* trabalhos não apresentam qualquer risco de causar uma perigosa explosão de inteligência. (Alguns paralelos históricos podem ser traçados com a trajetória da ideia da nanotecnologia; ver Drexler [2013].)

[244] Essas colaborações podem ser consideradas bem-sucedidas se considerarmos que elas alcançaram pelo menos alguns dos objetivos propostos inicialmente. O seu grau de sucesso num sentido mais amplo (levando em conta o custo-benefício e outros fatores) é mais difícil de determinar. No caso da Estação Espacial Internacional, por exemplo, ocorreram enormes atrasos e estouros de orçamento. Para detalhes sobre os problemas encontrados no projeto, ver NASA (2013). O projeto do Grande Colisor de Hádrons teve alguns contratempos significativos, mas isso pode ser devido à dificuldade inerente da tarefa. O Projeto Genoma Humano obteve sucesso ao final, mas parece ter recebido um incentivo por ter sido forçado a competir com o esforço da empresa privada de Craig Venter. Projetos patrocinados internacionalmente para obter energia a partir da fusão nuclear controlada falharam, apesar de investimentos massivos; mas, novamente, essas falhas podem ser atribuídas ao fato de que a tarefa era mais difícil do que se imaginava.

[245] *US Congress, Office of Technology Assessment* (1995).

[246] Hoffman (2009); Rhodes (2008).

[247] Rhodes (1986).

[248] A organização de criptógrafos da Marinha americana, OP-20-G, aparentemente ignorou um convite para obter total conhecimento sobre os métodos anti-Enigma dos britânicos e não informou o alto escalão dos Estados Unidos a respeito da oferta britânica de compartilhar seus segredos de criptografia (Burke, 2001). Isso deu aos líderes norte-americanos a impressão de que a Inglaterra estava ocultando informações importantes, o que foi uma razão de atrito ao longo da guerra. A Inglaterra compartilhou com o governo soviético parte do conhecimento que eles tinham colhido das decodificações das comunicações alemãs. Em particular, a Rússia foi alertada sobre os preparativos alemães para a Operação Barbarossa. Mas Stálin se recusou a acreditar no alerta, em parte porque os britânicos não revelaram como eles tinham conseguido tal informação.

[249] Por alguns anos, Russell parece ter defendido a ameaça de guerra nuclear para persuadir a Rússia a aceitar o plano Baruch; mais tarde, ele foi um forte proponente de um desarmamento nuclear mútuo (Russell e

Griffin, 2001). Alegou-se que John von Neumann acreditava que a guerra entre os Estados Unidos e a Rússia era inevitável, declarando que: “Se você perguntar por que não bombardeá-los [os russos] amanhã, eu pergunto: por que não bombardeá-los hoje? Se você propor hoje às cinco horas, eu direi: por que não a uma hora?”. (É possível que ele tenha feito essa notória declaração para reforçar suas credenciais anticomunistas com os gaviões do Departamento de Defesa dos Estados Unidos da era McCarthy. Mesmo que von Neumann tivesse sido o responsável pela política governamental dos Estados Unidos, é impossível saber se ele teria autorizado um primeiro ataque. Ver Blair [1957], p. 96.)

[250] Baratta (2004).

[251] Se a IA for controlada por um grupo de humanos, o problema pode se aplicar a esse grupo, embora seja possível que novas formas de comprometimento confiável a um acordo estejam disponíveis na época. Nesse caso, mesmo grupos humanos poderiam evitar o problema de uma potencial revolta interna e destituição por uma subcoalizão.

[252] Em que sentido a humanidade é uma espécie dominante sobre a Terra? Ecologicamente falando, entre os grandes animais (~50 kg), os humanos são os mais comuns, mas a biomassa seca total dos seres humanos (~100 bilhões kg) não é tão impressionante se comparada com a das formigas da família *Formicidae* (300 bilhões — 3.000 bilhões de kg). Os seres humanos e os organismos que lhes são úteis compõem apenas uma pequena parte (<0,001) de toda a biomassa global. Contudo, plantações e pastagens estão, hoje em dia, entre os maiores ecossistemas do planeta, cobrindo cerca de 35% das áreas não congeladas da superfície terrestre (Foley et al., 2007). E nós nos apropriamos de praticamente um quarto da produção primária líquida, de acordo com uma avaliação típica (Haberl et al., 2007), embora as estimativas variem de 3 a mais de 50%, dependendo principalmente de variações nas definições dos termos relevantes. Os seres humanos são responsáveis também por uma cobertura geográfica maior do que qualquer outra espécie animal e estão no topo da maioria das cadeias alimentares.

[253] Zalasiewicz et al. (2008).

[254] Ver a primeira nota deste capítulo.

[255] Estritamente falando, essa perspectiva pode não estar totalmente correta. A inteligência na espécie humana chega a aproximadamente zero (por exemplo, em embriões e pacientes em estado vegetativo permanente). Em termos qualitativos, a diferença máxima das habilidades cognitivas entre os indivíduos da espécie humana é, portanto, talvez até maior do que a diferença entre qualquer ser humano e uma superinteligência. Mas o argumento do texto continua valendo se consideramos o “ser humano” como um “adulto normal e funcional”.

[256] Gottfredson (2002). Ver também Carroll (1993) e Deary (2001).

[257] Ver Legg (2008). Resumidamente, Legg propõe a medição de um agente de aprendizagem por reforço de acordo com a sua expectativa de desempenho em todos os ambientes de recompensas possíveis, onde cada um desses ambientes receba um peso determinado pela sua complexidade de Kolmogorov. Explicaremos o que significa aprendizagem por reforço no capítulo 12. Ver também Dowe e Hernández-Orallo (2012) e Hibbard (2011).

[258] Em relação às pesquisas tecnológicas em áreas como a biotecnologia ou a nanotecnologia, uma superinteligência se sobressairia no projeto e na modelagem de novas estruturas. Supondo que a inventividade em projetos e modelagens não substitua a experimentação física, a vantagem de desempenho de uma superinteligência pode depender do seu grau de acesso ao aparato experimental necessário.

[259] Por exemplo, Drexler (1992, 2013).

[260] Uma IA de domínio restrito poderia, é claro, ter aplicações comerciais significativas, mas isso não significa que ela deteria o superpoder de produtividade econômica. Por exemplo, mesmo que uma IA de domínio restrito gerasse bilhões de dólares para o seu dono, esse valor ainda seria quatro ordens de grandeza menor do que o resto da economia mundial. Para que o sistema aumentasse direta e substancialmente a produção mundial, uma IA teria que ser capaz de realizar muitos tipos de trabalho; ou seja, ela precisaria ter competência em diversos domínios do conhecimento.

[261] Esse critério não exclui todos os cenários nos quais a IA falharia. Por exemplo, a IA poderia fazer uma aposta racional que tenha uma grande chance de fracasso. Nesse caso, entretanto, os critérios adotados poderiam levar em conta que (a) a IA deve fazer uma estimativa imparcial da baixa chance de sucesso de sua aposta e (b) não deve existir uma aposta melhor disponível para a IA que possa ser imaginada por nós, humanos atuais, e que não tenha sido levada em consideração por ela.

[262] Cf. Freitas (2000) e Vassar e Freitas (2006).

[263] Yudkowsky (2008a).

[264] Freitas (1980); Freitas e Merkle (2004, cap. 3); Armstrong e Sandberg (2013).

[265] Ver, por exemplo, Huffman e Pless (2003), Knill et al. (2002) e Drexler (1986).

[266] Consideramos aqui que não existam civilizações extraterrestres que possam se colocar em nosso caminho. Também estamos supondo que a hipótese da simulação é falsa. Ver Bostrom (2003a). Se qualquer uma dessas suposições estiver incorreta, podem haver riscos não antropogênicos significativos — que envolvam um agente inteligente de origem não humana. Ver também Bostrom (2003b, 2009c).

[267] Ao menos um *singleton* prudente que captasse a ideia de evolução poderia, em princípio, ter conduzido um programa eugênico por meio do qual lhe seria possível aumentar lentamente o seu nível de inteligência coletiva.

[268] Tetlock e Belkin (1996).

[269] De forma clara: colonizar e reprojetar uma grande parcela do universo acessível não está atualmente ao nosso alcance *direto*. A colonização intergaláctica está muito além de nossa tecnologia atual. O ponto aqui é que nós poderíamos, em princípio, usar nossas capacidades atuais para desenvolver as capacidades adicionais que seriam necessárias, colocando, assim, essa empreitada ao nosso alcance *indireto*. É também igualmente verdade que a humanidade não é atualmente um *singleton* e que não sabemos se enfrentaremos ou não a oposição inteligente de algum poder externo caso iniciemos um processo de modificação do universo acessível. Entretanto, para alcançar o patamar de sustentabilidade de um *singleton* prudente, é suficiente possuir um conjunto de capacidades tal que, se um *singleton* prudente que não estivesse enfrentando oposição inteligente o possuísse, então a colonização e a modificação do universo acessível estariam ao seu alcance indireto.

[270] Esses números são baseados na estimativa WMAP de uma densidade bariônica de 9,9×10-30 g/cm³ e considera que 90% da massa do universo é formada por gases intergalácticos, que 15% da massa galáctica é constituída de estrelas (sendo cerca de 80% de matéria bariônica) e que a massa de uma estrela média é 0,7 da massa solar (Read e Trentham, 2005; Carrol e Ostlie, 2007).

[271] Armstrong e Sandberg (2013).

[272] Mesmo a 100% de *c* (o que é inatingível para objetos com massa em repouso superior a zero), o número de galáxias alcançáveis é, aproximadamente, apenas 6×10^9. (Cf. Gott et al. [2005] e Heyl [2005].) Isso se considerarmos que nossa atual compreensão da física relevante esteja correta. Dificilmente poderíamos ser muito confiantes em relação a qualquer limite superior, uma vez que é ao menos concebível que uma civilização superinteligente possa estender seu alcance de alguma forma que nós imaginamos fisicamente impossível (por exemplo, com a construção de máquinas do tempo, por meio da criação de novos universos inflacionários ou por outros meios ainda não imaginados).

[273] O número de planetas habitáveis por estrela é atualmente incerto, então se trata meramente de uma estimativa grosseira. Traub (2012) prevê que um terço das estrelas das classes espectrais F, G ou K tenham ao menos um planeta terrestre na zona habitável; ver também Clavin (2012). As estrelas FGK representam cerca de 22,7% das estrelas vizinhas ao nosso sistema solar, o que sugere que 7,6% das estrelas possuem planetas potencialmente apropriados. Além disso, talvez existam planetas habitáveis orbitando em torno de estrelas M (Gilster, 2012), que são mais numerosas. Ver também Robles et al. (2008). Não seria necessário expor corpos humanos aos rigores das viagens intergalácticas. IAs poderiam supervisionar o processo de colonização. O *Homo sapiens* poderia ser levado junto como fonte de "informação", que poderia ser posteriormente utilizada pelas IAs para a criação de indivíduos de nossa espécie. Por exemplo, informação genética poderia ser sintetizada na forma de DNA e uma primeira geração de humanos seria incubada, criada e educada por IAs guardiãs que assumissem uma aparência antropomórfica.

[274] O'Neill (1974).

[275] Dyson (1960) diz que sua ideia foi inspirada nos trabalhos do escritor de ficção científica Olaf Stapledon (1937), que, por sua vez parece ter se inspirado em J.D. Bernal (Dyson 1979, p. 211).

[276] Segundo o princípio de Landauer, há uma quantidade mínima de energia necessária para modificar um bit de dados, conhecida como o limite de Landauer, igual a $kT \ln 2$, onde k é a constante de Boltzmann ($1{,}38\times10^{-23}$ J/K) e T é a temperatura. Se considerarmos que os circuitos são mantidos a cerca de 300 K, então 10^{26} watts permitem que apaguemos aproximadamente 10^{47} bits por segundo. (A respeito da eficiência possível a dispositivos computacionais nanomecânicos, ver Drexler [1992]. Ver também Bradbury [1999]; Sandberg [1999]; Cirkovic [2004]. Os fundamentos do princípio de Landauer são um tanto controversos; ver, por exemplo, Norton [2011].)

[277] As estrelas variam em seus níveis de produção de energia, mas o Sol é o uma estrela de sequência principal relativamente padrão.

[278] Uma análise mais detalhada poderia considerar mais atentamente os tipos de operações que nos interessa. O número de operações *seriais* que podem ser realizadas é bastante limitado, uma vez que um computador serial rápido precisa ser pequeno para conseguir minimizar atrasos de comunicação entre suas

partes. Há também uma limitação quanto ao número de bits que podem ser armazenados e, como vimos, em relação ao número de passos computacionais irreversíveis (que envolvem a eliminação de informação) que podem ser realizados.

[279] Às vezes pode ser útil falar de duas IAs, ambas dotadas de um determinado superpoder. Em um entendimento amplo do termo, é possível conceber um superpoder como algo que um agente possua em relação a algum campo específico de atuação — em nosso exemplo, talvez um campo que inclua toda a civilização humana, mas exclua a outra IA.

[280] Isso, é claro, não significa dizer que diferenças que pareçam visualmente pequenas não possam ser funcionalmente profundas.

[281] Yudkowsky (2008a, p. 310).

[282] David Hume, filósofo iluminista escocês, acreditava que crenças, por si mesmas (por exemplo, sobre o que seria uma coisa boa a se fazer), não podem motivar ações: algum desejo é necessário. Isso sustentaria a tese da ortogonalidade por meio da eliminação de uma possível objeção em relação ela: que um nível de inteligência suficiente pode resultar na aquisição de certas crenças, o que, então, necessariamente produziria determinadas motivações. No entanto, apesar da tese da ortogonalidade ter amparo na teoria de motivação de Hume, ela não a pressupõe. Em particular, a tese da ortogonalidade não pressupõe que crenças, por si mesmas, não possam nunca motivar uma ação. Seria suficiente admitir, por exemplo, que um agente — seja qual for sua inteligência — pode ser motivado a seguir qualquer curso de ação caso ele tenha certos desejos suficientemente fortes para anular o efeito das crenças. Outro cenário em que a tese da ortogonalidade poderia ser verdadeira mesmo que a teoria da motivação de Hume seja falsa seria possível se uma inteligência arbitrariamente alta não implicar na aquisição de qualquer uma dessas crenças que são (supostamente) motivadoras por si mesmas. Um terceiro modo através do qual seria possível que a tese da ortogonalidade fosse verdadeira mesmo que a teoria de Hume resultasse falsa se daria com a possibilidade de construir um agente (ou, de forma mais neutra, um "processo de otimização") dotado de uma inteligência arbitrariamente alta, mas com uma constituição tão alienígena que não teria nenhum análogo funcional ao que nós, humanos, chamamos de "crenças" e "desejos". (Para algumas tentativas recentes de defesa da teoria de motivação de Hume, ver Smith [1987], Lewis [1988] e Sinhababu [2009].)

[283] Por exemplo, Derek Parfit argumentou que certas preferências básicas seriam irracionais, tal como o caso de um agente normal que apresente "indiferença-à-terça-feira-futura": um certo hedonista se importa bastante com a qualidade de suas experiências futuras. Com uma exceção, ele se importa igualmente com todas as partes do seu futuro. A exceção é que ele possui "indiferença-à-terça-feira-futura". Em todas as terças-feiras ele se importa normalmente com o que pode estar lhe acontecendo. Mas ele nunca se importa com as possíveis dores ou prazeres de uma terça-feira futura [...]. Essa indiferença é um simples fato. Quando ele planeja seu futuro, é simplesmente verdade que ele sempre prefere a perspectiva de um grande sofrimento numa terça-feira a um simples desconforto em qualquer outro dia. (Parfit [1986, p. 123-4]; ver também Parfit [2001]). Para os nossos propósitos, não é necessário nos posicionar em relação ao fato de Parfit estar correto ou não a respeito da irracionalidade do agente, contanto que admitamos que ele não é necessariamente não inteligente no sentido instrumental explicado no texto. O agente de Parfit poderia ter uma racionalidade instrumental impecável e, portanto, grande inteligência, mesmo que lhe falte certa sensibilidade à "razão objetiva" que poderia ser necessária para um agente completamente racional. Sendo assim, esse tipo de exemplo não invalida a tese da ortogonalidade.

[284] Mesmo que haja fatos morais objetivos que poderiam ser compreendidos por qualquer agente completamente racional, e mesmo que tais fatos morais fossem de alguma forma intrinsecamente motivadores (de maneira que qualquer pessoa com total capacidade de compreendê-los seja forçosamente motivada a agir de acordo com eles), não necessariamente poderíamos invalidar a tese da ortogonalidade. A tese ainda poderia ser verdadeira se um agente pudesse ter uma racionalidade *instrumental* impecável, mesmo que não apresentasse algumas outras faculdades constitutivas de racionalidade propriamente dita, ou alguma faculdade exigida para a total compreensão dos fatos morais objetivos. (Um agente poderia também ser extremamente inteligente, até mesmo superinteligente, ainda que ele não tivesse uma racionalidade instrumental total em todas as áreas do conhecimento.)

[285] Para mais informações sobre a tese da ortogonalidade, ver Bostrom (2012) e Armstrong (2013).

[286] Sandberg e Bostrom (2008).

[287] Stephen Omohundro escreveu dois textos pioneiros sobre esse tópico (Omohundro, 2007, 2008). Omohundro defende que todos os sistemas avançados de IA são suscetíveis de apresentar um número de "impulsos básicos", o que para ele quer dizer "tendências que estarão presentes, a menos que sejam explicitamente combatidas". O termo "impulso de IA" possui a vantagem de ser curto e sugestivo, mas tem a desvantagem de sugerir que os objetivos instrumentais aos quais se refere influenciam a capacidade da IA de tomar decisões, do mesmo modo que os impulsos psicológicos influenciam nas tomadas de decisão dos

humanos, ou seja, através de uma espécie de “puxão fenomenológico” em nosso ego, ao qual nossa força de vontade pode ocasionalmente oferecer resistência. Essa conotação não é útil. Ninguém normalmente diria que um ser humano típico possui um “impulso” para preencher sua declaração de impostos, ainda que preenchê-la possa ser um objetivo instrumental convergente para os humanos das sociedades contemporâneas (um objetivo cuja realização evita problemas que nos impediriam de realizar muitos de nossos objetivos finais). Nossa abordagem aqui difere daquela de Omohundro em alguns outros aspectos mais substanciais, embora a ideia subjacente seja a mesma. (Ver também Chalmers [2010] e Omohundro [2012].)

[288] Chislenko (1997).

[289] Ver também Shulman (2010b).

[290] Um agente poderia mudar também a *representação* de seus objetivos se isso for capaz de modificar sua ontologia, de modo a transpor a representação antiga para a nova ontologia; cf. de Blanc (2011). Outro tipo de fator que pode fazer um *teórico da decisão baseado em evidências* tomar ações variadas, incluindo a mudança de seus objetivos finais, é a relevância das evidências a favor dessa decisão. Por exemplo, um agente que siga a teoria da decisão baseada em evidências pode acreditar que haja outros agentes como ele no universo e que suas próprias ações fornecerão alguma evidência sobre como esses outros agentes vão agir. Esse agente pode, assim, optar por adotar um objetivo final que seja altruístico em relação aos outros agentes ligados pela evidência, partindo da possibilidade de que essa escolha fornecerá evidências de que esses outros agentes terão escolhido agir da mesma forma. Entretanto, um resultado semelhante pode ser obtido sem a mudança dos objetivos finais, e então o agente escolheria a cada instante agir *como se* esses fossem seus objetivos finais.

[291] Uma extensiva literatura de psicologia explora a formação adaptativa de preferências. Ver, por exemplo, Forgas et al. (2010).

[292] Em modelos formais, o valor da informação é quantificado como a diferença entre o valor que se espera obter por meio das decisões ótimas feitas com base nas informações e o valor que se espera obter com as decisões ótimas feitas sem base nas informações. (Ver, por exemplo, Russell e Norvig [2010].) A partir dessa definição, podemos concluir que o valor da informação nunca é negativo. E, também, que qualquer informação que alguém possua, mas que nunca afetará qualquer decisão que essa pessoa venha a tomar, tem valor zero para ela. Entretanto, esse tipo de modelo toma como verdadeiras diversas idealizações que são geralmente inválidas no mundo real — por exemplo, que o conhecimento não tem valor final (o que significa que o conhecimento possui apenas valor instrumental, não apresentando valor por si próprio) e que agentes não são transparentes com outros agentes.

[293] Por exemplo, Hájek (2009).

[294] Essa estratégia é exemplificada pela larva da ascídia (seringa-do-mar), que nada até encontrar uma rocha adequada, à qual ela se fixa permanentemente. Cimentada em seu lugar, a larva tem menos necessidade de processamento de informações complexas, motivo pelo qual digere parte do seu próprio cérebro (seu gânglio cerebral). O mesmo fenômeno pode ser observado no caso de alguns acadêmicos que tenham garantido sua estabilidade.

[295] Bostrom (2012).

[296] Bostrom (2006c).

[297] Poderíamos reverter a questão e examinar as razões possíveis para que um *singleton* superinteligente *não* desenvolva algumas capacidades tecnológicas. Elas incluem: (a) o *singleton* prevê que a capacidade não terá utilidade; (b) o custo de desenvolvimento é muito grande quando comparado com sua utilidade prevista (por exemplo, se a tecnologia nunca fosse adequada para alcançar qualquer um dos objetivos do *singleton*, ou se o *singleto*n tiver uma taxa de desconto muito alta que desencoraje investimentos fortemente); (c) o *singleton* tem algum valor final que requer a abstenção de certas vias de desenvolvimento tecnológico; (d) se o *singleton* não tiver certeza de que permanecerá estável, ele pode preferir se abster de desenvolver tecnologias que poderiam ameaçar sua estabilidade interna ou que fariam com que as consequências de dissolução sejam piores (por exemplo, um governo mundial pode escolher não desenvolver tecnologias que facilitem rebeliões, mesmo que tenham um bom uso para elas, nem desenvolver tecnologias que facilitassem a produção de armas de destruição em massa que poderiam levar à devastação caso esse governo mundial se dissolvesse); (e) similarmente, o *singleton* poderia ter feito alguma espécie de acordo estratégico para o não desenvolvimento de determinada tecnologia, o qual se manteria válido mesmo se a partir de algum momento fosse conveniente desenvolvê-la. (Note, entretanto, que algumas razões *atuais* para o desenvolvimento tecnológico *não* se aplicariam a um *singleton*: por exemplo, razões ligadas a uma corrida armamentista.)

[298] Suponha que um agente desconte recursos obtidos no futuro a uma taxa exponencial e que, em virtude da limitação da velocidade da luz, o agente somente possa aumentar a quantidade de recursos que possui a uma razão polinomial. Isso significa que em algum momento o agente não veria mais utilidade em continuar expandindo sua aquisição de recursos? Não, pois embora o valor atual dos recursos obtidos em tempos futuros

seja uma assíntota tendendo a zero quanto mais à frente olhamos, *esse também seria o custo de sua obtenção no presente.* O custo atual do lançamento de outra sonda de von Neumann daqui a 100 milhões de anos (possivelmente utilizando alguns recursos adquiridos pouco tempo antes) seria reduzido pelo mesmo fator de desconto que reduziria os valores atuais dos recursos futuros adquiridos por essa sonda extra (excluindo-se um fator constante).

[299] Enquanto o volume alcançado pelas sondas de colonização durante um certo período poderia ser aproximadamente esférico e se expandir a uma taxa proporcional ao quadrado do tempo transcorrido desde o lançamento da primeira sonda ($\sim t^2$), a quantidade de recursos contida nesse volume seguirá um padrão de crescimento menos regular, uma vez que a distribuição de recursos não é homogênea e varia em diversas escalas. Inicialmente, a taxa de crescimento pode ser $\sim t^2$ enquanto o planeta natal é colonizado; a seguir, a taxa de crescimento pode dar saltos à medida que planetas e sistemas solares próximos forem colonizados; então, conforme o volume em formato de disco da Via Láctea for sendo preenchido, a taxa de crescimento pode se equilibrar, tornando-se aproximadamente proporcional a t; depois disso, a taxa de crescimento pode novamente apresentar saltos conforme galáxias próximas forem sendo colonizadas; em seguida, a taxa de crescimento novamente poderia se aproximar de $\sim t^2$, conforme a expansão procedesse em uma escala na qual a distribuição das galáxias seja mais ou menos homogênea; então, outro período de saltos de crescimento seguidos por um suave crescimento de $\sim t^2$ conforme superaglomerados de galáxias sejam colonizados; até que finalmente a taxa de crescimento entre em declínio, alcançando zero quando a velocidade de expansão do universo aumentar de tal modo que futuras colonizações se tornem impossíveis.

[300] O argumento de simulação pode ser de particular importância dentro desse contexto. Um agente superinteligente pode atribuir uma probabilidade significativa a hipóteses de acordo com as quais ele viva em uma simulação de computador e sua sequência de percepções seja gerada por outra superinteligência, o que pode criar várias razões instrumentais convergentes, a depender das suposições do agente a respeito dos tipos de simulações em que ele provavelmente se encontra. Cf. Bostrom (2003a).

[301] Descobrir as leis básicas da física e outros fatores fundamentais relacionados ao mundo é um objetivo instrumental convergente. Podemos classificá-lo aqui como uma “melhoria cognitiva”, embora também pudesse ser derivado do objetivo de “aperfeiçoamento tecnológico” (uma vez que novos fenômenos físicos possibilitariam novas tecnologias).

[302] Existem alguns riscos existenciais adicionais em cenários nos quais a humanidade sobreviva em um estado de acentuado subaproveitamento ou em que uma grande parte do nosso potencial para um desenvolvimento desejável seja irreversivelmente desperdiçado. Além disso, podem haver riscos existenciais associados ao processo que conduziria a uma potencial explosão de inteligência, a qual poderia surgir, por exemplo, num contexto de guerras entre nações que estivessem competindo pelo desenvolvimento da primeira superinteligência.

[303] Há um momento de vulnerabilidade significativo quando a IA se dá conta da necessidade de ocultar suas capacidades (uma situação que podemos chamar de *concepção da mentira*). Essa percepção inicial não seria deliberadamente ocultada no momento em que ocorresse. Mas uma vez que essa percepção ocorra, a IA poderia tentar esconder rapidamente o fato de que tal percepção ocorreu, estabelecendo simultaneamente alguma dinâmica interna oculta (talvez disfarçada em algum processo inócuo que se misture com todos os outros processos complicados que ocorrem em sua mente) que possibilitaria que ela continuasse a planejar sua estratégia de longo prazo em segredo.

[304] Mesmo hackers humanos conseguem escrever programas pequenos e aparentemente inócuos capazes de fazer coisas completamente inesperadas. (Considere, por exemplo, alguns dos vencedores da Competição Internacional de Código C Ofuscado.)

[305] A observação de que algumas medidas de controle de IAs poderiam dar a impressão de estar funcionando dentro de um contexto fixo e ainda assim falharem catastroficamente quando esse contexto se altera também é enfatizada por Eliezer Yudkowsky; ver, por exemplo, Yudkowsky (2008a).

[306] O termo parece ter sido cunhado pelo escritor de ficção científica Larry Niven (1973), mas é baseado em experimentos de recompensa através da estimulação do cérebro no mundo real; cf. Olds e Milner (1954) e Oshima e Katayama (2010). Ver também Ring e Orseau (2011).

[307] Bostrom (1997).

[308] É possível que existam algumas implementações possíveis de um mecanismo de aprendizagem por reforço que conduziriam, uma vez que a IA descobrisse a solução de *wireheading*, a uma incapacitação segura em vez de uma profusão de infraestrutura. A questão é que isso poderia facilmente dar errado e falhar por motivos inesperados.

[309] Isso foi sugerido por Marvin Minsky (ver Russell e Norvig [2010, p. 1039]).

[310] A questão a respeito de quais tipos de mentes digitais poderiam ser conscientes no sentido de possuir uma experiência fenomenal subjetiva, ou "qualia", no jargão da filosofia, é importante para esse ponto (embora seja irrelevante para muitas outras partes deste livro). Uma questão em aberto é o quão difícil seria estimar precisamente como um ser semelhante ao humano se comportaria em diversas circunstâncias sem que fosse necessário simular seu cérebro com detalhes suficientes para que a simulação chegasse a ser consciente. Uma outra questão é se existem ou não algoritmos que sejam úteis, de uma forma geral, a uma superinteligência, como técnicas de aprendizagem por reforço, de maneira que a implementação desses algoritmos fosse capaz de gerar qualia. Ainda que julguemos bem pequena a probabilidade de que qualquer uma dessas sub-rotinas seja consciente, o número de instanciações poderia ser tão grande que até um pequeno risco de que elas possam experimentar sofrimento deve ter um peso significativo em nosso cálculo moral. Ver também Metzinger (2003, cap. 8).

[311] Bostrom (2002a, 2003a); Elga (2004).

[312] O termo "agência" é utilizado aqui para se referir à capacidade de atuação autônoma de um indivíduo, quer ele seja humano ou IA. [NT]

[313] Ver, por exemplo, Laffont e Martimort (2002).

[314] Suponha que a maioria dos eleitores deseja que seu país construa algum tipo específico de superinteligência. Eles elegem, então, um candidato que promete fazer exatamente isso, mas eles acham difícil assegurar que ele, uma vez no poder, cumprirá sua promessa de campanha e finalizará o projeto da maneira que os eleitores desejavam. Suponhamos que ele viesse a ser fiel à sua promessa, instruindo o governo a contratar um consórcio de empresas ou instituições acadêmicas para a realização do trabalho; mas, novamente, haveria problemas de agência; os burocratas no departamento do governo poderiam ter suas próprias opiniões a respeito do que deveria ser feito, executando o projeto de modo que respeitasse a forma, e não o espírito, das instruções do líder. Mesmo que o departamento do governo cumprisse à risca seu papel, os parceiros científicos contratados poderiam ter intenções próprias. O problema se repete em diversos níveis. O diretor de um dos laboratórios participantes passaria a noite acordado se preocupando com o fato de que um técnico poderia introduzir algum elemento não sancionado no projeto — imaginando o dr. T.R. Eason entrando furtivamente em seu escritório tarde da noite, acessando a base de código do projeto, reescrevendo uma parte do sistema de objetivo da IA embrionária e, onde supostamente se lia "sirva a humanidade", ele escreveria "sirva o dr. T.R. Eason".

[315] Todavia, até mesmo para o desenvolvimento de uma superinteligência, deveria haver testes comportamentais — como um elemento auxiliar dentro de uma bateria de medidas de segurança mais amplas. Se uma IA se comportasse indevidamente na sua fase de desenvolvimento, algo estaria claramente errado — embora, é importante lembrar, o inverso não seja verdadeiro.

[316] Em um conhecido episódio ocorrido no ano de 1975, Steven Dompier escreveu um programa para o Altair 8800, o qual tirou vantagem desse efeito (e da ausência de proteção ao redor da caixa do microcomputador). Durante a execução do programa, ondas eletromagnéticas eram emitidas e produziam música quando alguém segurava um rádio transistorizado próximo ao computador (Driscoll, 2012). O jovem Bill Gates, que participou de uma demonstração, disse ter ficado impressionado e perplexo com o feito (Gates, 1975). De todo modo, há planos para se projetar chips com capacidades de wi-fi embutidas (Greene, 2012).

[317] Não é insignificante o fato de que, se tivéssemos tido a oportunidade de agir de acordo com essa convicção equivocada, ela poderia ter resultado na destruição de nossa herança cósmica como um todo. Talvez faça sentido argumentar em favor do seguinte princípio: se no passado tivesse existido a certeza de que em N ocasiões um sistema melhorou o suficiente para se tornar seguro e em cada uma dessas ocasiões fossem revelados equívocos a esse respeito, então, na ocasião seguinte, não seria correto designar valores maiores do que $1/(N + 1)$ para a probabilidade de segurança do sistema.

[318] Em um experimento informal, um humano inteligente fez o papel da IA e outro indivíduo atuou como o guardião que estava instruído a não permitir que a IA saísse do confinamento. A IA poderia se comunicar com o guardião apenas por meio de textos e tinha duas horas para persuadi-lo a deixá-la sair. Em três dos cinco testes, nos quais indivíduos diferentes fizeram o papel do guardião, a IA escapou (Yudkowsky, 2002). Uma tarefa na qual um humano foi bem-sucedido certamente não seria difícil para uma superinteligência. (O inverso, é claro, não é verdadeiro. Mesmo que a tarefa dada a uma superinteligência real fosse mais difícil — talvez os guardiões estivessem mais firmemente motivados em não liberar a IA do que os indivíduos que fizeram o papel de guardião no experimento —, a superinteligência poderia, ainda assim, ser bem-sucedida em casos nos quais um humano falharia.)

[319] Não se deve exagerar a quantidade mínima de segurança que poderia ser obtida dessa maneira. Imagens mentais podem substituir imagens em um monitor gráfico. Considere o impacto que os livros podem causar nas pessoas — e livros não são nem ao menos interativos.

[320] Ver também Chalmers (2010). Seria um erro deduzir, com base nisso, que não há *nenhuma* utilidade possível na construção de um sistema que nunca seria observado por qualquer entidade externa. Seria possível atribuir um valor final ao que acontece dentro de tal sistema. Além disso, outras pessoas também poderiam ter preferências em relação ao que acontece dentro desse sistema, sendo, assim, influenciadas por sua criação ou pela promessa de sua criação. O conhecimento da existência de certos tipos de sistemas isolados (que contêm observadores) pode também induzir os observadores externos à incerteza antrópica, o que poderia influenciar seus comportamentos.

[321] Alguém poderia se perguntar por que a integração social é considerada uma forma de controle de capacidade. Ela não deveria, em vez disso, ser classificada como um método de seleção de motivação, uma vez que procura influenciar o comportamento de um sistema por meio de incentivos? Examinaremos atentamente a seleção de motivação neste capítulo; mas, respondendo à questão, nós estamos definindo a seleção de motivação como um grupo de métodos de controle que trabalham por meio da seleção ou formatação dos objetivos finais do sistema — objetivos que são buscados como um fim em si próprios, e não por razões instrumentais. A integração social não visa influenciar os objetivos finais do sistema, logo não é uma seleção de motivação. Ao contrário, a integração social objetiva limitar as capacidades efetivas do sistema, buscando fazer com que este seja incapaz de conseguir um certo conjunto de resultados — através dos quais poderia optar por desistir sem sofrer as penalidades associadas (retaliação e perda dos ganhos com a colaboração). A esperança é que, limitando-se os resultados que estaria apto a obter, o sistema pudesse compreender que a maneira mais efetiva de atingir seus objetivos finais seria por meio da cooperação.

[322] Para o artigo original, ver Bostrom (2003a). Ver também Elga (2004).

[323] Shulman (2010a).

[324] A realidade concreta, presumidamente, contém mais recursos computacionais do que uma realidade simulada, uma vez que qualquer processo computacional que ocorra em uma simulação ocorrerá simultaneamente no computador que está executando essa simulação. A realidade concreta pode também conter uma riqueza de outros recursos físicos que poderiam ser de difícil acesso para os agentes simulados — agentes que existem apenas à mercê de simuladores poderosos que podem ter outras finalidades em mente para esses recursos. (É claro que a inferência aqui não é válida de forma estritamente dedutível: em princípio, poderia ser verdade que os universos nos quais as simulações são executadas contenham muito mais recursos, de forma que as civilizações simuladas tivessem, em média, acesso a mais recursos do que as civilizações não simuladas, embora cada civilização não simulada que executasse simulações teria mais recursos do que todas as civilizações que ela simula combinadas.)

[325] L&PM, 1997. Trad. Millôr Fernandes. [NE]

[326] Há várias outras considerações esotéricas que poderiam pesar sobre essa questão, cujas implicações ainda não foram analisadas a fundo. Essas considerações podem, no fim das contas, ser importantíssimas para o desenvolvimento de uma abordagem completa para lidar com a perspectiva de uma explosão de inteligência. Entretanto, parece improvável que conseguiremos imaginar a importância prática de tais argumentos esotéricos, a menos que tenhamos primeiramente progredido em relação aos tipos mais mundanos de considerações, as quais são o assunto de boa parte deste livro.

[327] Essa abordagem poderia ser mais promissora no caso de uma emulação que acreditemos ter motivações antropomórficas.

[328] Devo essa ideia a Carl Shulman.

[329] Criar uma criptografia capaz de resistir a um decifrador de códigos superinteligente não é um desafio trivial. Por exemplo, traços de números aleatórios podem ser deixados no cérebro de algum observador ou em alguma microestrutura de um gerador aleatório, de onde a superinteligência poderia tirá-los; ou, se números pseudoaleatórios forem usados, a superinteligência poderá adivinhá-los ou descobrir a semente a partir da qual eles foram gerados. Além disso, a superinteligência poderia construir grandes computadores quânticos, ou até mesmo descobrir fenômenos físicos desconhecidos que poderiam ser utilizados para construir novos tipos de computadores.

[330] A IA poderia modificar sua própria estrutura de modo a *acreditar* que ela teria recebido esses prêmios, mas isso não caracterizaria um *wireheading* se ela fosse projetada para desejar os prêmios (em vez de desejar permanecer em um estado no qual ela tivesse certas crenças a respeito deles).

[331] Cf., por exemplo, Quine e Ullian (1978).

[332] O que uma IA poderia investigar levando em consideração as características de desempenho de várias funcionalidades básicas computacionais, tais como o tamanho e capacidade de vários barramentos de dados, o tempo necessário para acessar diferentes partes da memória, a incidência de inversões aleatórias de bits, e assim por diante.

[333] Talvez a probabilidade *a priori* pudesse ser uma aproximação computável da probabilidade *a priori* de Solomonoff, que atribui probabilidade para mundos possíveis com base em suas complexidades algorítmicas. Ver Li e Vitányi (2008).

[334] Asimov (1942). A essas três leis foi adicionada, mais tarde, a "Lei Zero": "(0) Um robô não pode causar mal à humanidade ou, por inação, permitir que a humanidade sofra algum mal" (Asimov, 1985).

[335] Cf. Gunn (1982).

[336] Russell (1986, 161f).

[337] De modo semelhante, embora alguns filósofos tenham passado suas carreiras tentando formular cuidadosamente sistemas deontológicos, o surgimento de novos casos e novas consequências esporadicamente tornam necessárias revisões. Por exemplo, a filosofia moral deontológica tem, em anos recentes, se revigorado por conta da descoberta de uma nova classe fértil de experimentos de pensamento filosóficos, "dilemas do bonde", que revelam muitas interações sutis entre as nossas intuições em relação à importância moral da distinção entre atos e omissões, a distinção entre consequências planejadas e não planejadas, e outros temas semelhantes; ver, por exemplo, Kamm (2007).

[338] Armstrong (2010).

[339] Como regra geral, se planejamos usar múltiplos mecanismos de segurança para conter uma IA, pode ser prudente trabalhar com cada um desses mecanismos *como se* fossem planejados para ser o único mecanismo de segurança existente e *como se* fossem, portanto, autossuficientes. Se alguém coloca um balde rachado dentro de outro balde rachado, a água ainda assim continua vazando.

[340] Uma variação dessa mesma ideia consiste na construção da IA de modo que ela possa ser continuamente motivada a agir de acordo com suas melhores suposições em relação ao que seria o padrão definido implicitamente. Nessa configuração, o objetivo final da IA seria sempre o de agir de acordo com o padrão definido implicitamente, e ela buscaria investigar qual seria esse padrão apenas por razões instrumentais.

[341] Esses termos são obviamente antropomórficos e não devem ser considerados seriamente como analogias. Eles são apenas rótulos para uma espécie de *prima facie* de diferentes conceitos para possíveis sistemas que alguém poderia pensar em construir.

[342] Em resposta a uma pergunta sobre o resultado das próximas eleições, não gostaríamos de receber uma lista completa da posição projetada e dos vetores de momento de partículas próximas.

[343] Indexadas de acordo com um conjunto particular de instruções definido para uma máquina em particular.

[344] Kuhn (1962); de Blanc (2011).

[345] Seria mais difícil aplicar tal "método do consenso" a gênios e soberanos, pois geralmente há inúmeras sequências de ações básicas (tal como enviar padrões particulares de sinais elétricos para os atuadores do sistema) que seriam quase igualmente efetivas em alcançar um dado objetivo; de onde agentes ligeiramente diferentes poderiam legitimamente escolher ações ligeiramente diferentes, resultando na incapacidade de se chegar a um consenso. Em contraste, com perguntas formuladas apropriadamente, haveria, geralmente, um número pequeno de respostas adequadas (como "sim" ou "não"). (Sobre o conceito do ponto de Schelling, também conhecido como "ponto focal", ver Schelling [1980].)

[346] A economia mundial não seria, em alguns aspectos, análoga a um gênio fraco, embora seja um gênio que cobra por seus serviços? Uma economia muitíssimo maior, que poderá se desenvolver no futuro, se aproximaria de um gênio com uma superinteligência coletiva. Um importante aspecto que *difere* a economia atual de um gênio é que, embora seja possível fornecer um comando (que tem um preço) para que a economia entregue uma pizza na sua porta, não é possível dar o comando para que ela entregue paz. Isso não acontece porque a economia é insuficientemente poderosa, mas porque ela é insuficientemente coordenada. Nesse aspecto, a economia se parece mais com uma *assembleia* de gênios servindo mestres diferentes (com objetivos distintos) do que com um único gênio ou qualquer outro tipo de agente unificado. Aumentar o poder absoluto da economia, tornando cada um desses gênios mais poderosos ou adicionando mais gênios ao grupo, não faria com que a economia se tornasse mais capaz de cumprir o comando de entregar paz. Para funcionar como um gênio superinteligente, a economia não deveria apenas aumentar sua capacidade de produzir bens e serviços a um custo baixo (incluindo aqueles que demandam tecnologia radicalmente nova), mas também teria que aprimorar sua capacidade de resolver problemas de coordenação global.

[347] Caso o gênio fosse de algum modo incapaz de não obedecer a um comando subsequente — e de alguma maneira incapaz de se reprogramar para se livrar dessa suscetibilidade —, então ele poderia agir de forma a evitar que qualquer novo comando fosse emitido.

[348] Mesmo um oráculo que esteja limitado a fornecer respostas do tipo sim ou não poderia facilitar a busca por uma IA do tipo gênio ou soberano, ou mesmo ser usado diretamente como um componente de tal IA. O oráculo poderia também ser usado para gerar o próprio código dessa IA caso um número suficientemente

grande de perguntas pudessem ser feitas. Essa série de perguntas poderia seguir mais ou menos o seguinte roteiro: “Na versão binária do código da primeira IA em que você pensou para um gênio, o *n*-ésimo símbolo é zero?”.

[349] É possível imaginar um oráculo ou um gênio mais complexos, que aceitem perguntas ou comandos apenas se eles foram feitos por uma determinada autoridade, embora isso não exclua a possibilidade de que essa autoridade se corrompa ou seja chantageada por terceiros.

[350] John Rawls, um renomado filósofo político do século XX, tornou famosa a analogia do véu da ignorância como um meio de caracterizar os tipos de preferências que deveriam ser levadas em consideração na formulação de um contrato social. Rawls sugere que devemos imaginar que estamos cobertos por um véu de ignorância enquanto escolhemos um contrato social, o que nos impede de saber que tipo de pessoa nós seremos e que tipo de papel social ocuparemos. A ideia aqui é que, em tal situação, deveríamos pensar que tipo de sociedade seria geralmente mais justa e desejável sem levar em consideração nossos interesses egoístas e oportunistas que poderiam, de outro modo, fazer com que preferíssemos uma ordem social na qual usufruiríamos de privilégios injustos. Ver Rawls (1971).

[351] Karnofsky (2012).

[352] Uma possível exceção seriam os softwares ligados a atuadores suficientemente poderosos, como softwares de sistemas de alarme que estejam conectados diretamente a ogivas nucleares ou oficiais humanos autorizados a lançar um ataque nuclear. O mal funcionamento desses softwares pode resultar em situações de alto risco. Isso aconteceu pelo menos duas vezes na história. Em 9 de novembro de 1979, um problema nos computadores levou o Centro de Defesa Aeroespacial Norte-Americano (NORAD, na sigla em inglês) a emitir um falso alerta sobre um possível ataque nuclear massivo da União Soviética contra os Estados Unidos. Os Estados Unidos prepararam medidas emergenciais de retaliação antes que os alertas nos sistemas de radares apresentassem dados mostrando que tal ataque não havia sido lançado (McLean e Stewart, 1979). Em 26 de setembro de 1983, um mal funcionamento no sistema soviético de alertas nucleares, conhecido como Oko, resultou na emissão de um alerta sobre um ataque de mísseis dos Estados Unidos. Tal alerta foi corretamente identificado pelo oficial responsável pelo centro de comandos, Stanislav Petrov, como um alarme falso: essa decisão foi considerada responsável por ter evitado uma guerra termonuclear (Lebedev, 2004). Ao que tudo indica, essa guerra não teria levado à extinção da humanidade, mesmo que fosse utilizado o arsenal combinado de todas as potências nucleares disponível no auge da Guerra Fria, mas teria arruinado a civilização e causado mortes e sofrimento inimagináveis (Gaddis, 1982; Parrington, 1997). Mas estoques maiores podem ser acumulados em uma corrida armamentista futura, ou armas ainda mais letais podem ser desenvolvidas, e nossos modelos para os impactos de um Armagedom nuclear (particularmente a respeito da severidade de um inverno nuclear resultante) podem estar errados.

[353] Essa abordagem pode se encaixar na categoria de um método de controle de especificação direta baseado em regras.

[354] A situação seria essencialmente a mesma se o critério de solução estabelecer uma *medida* de adequação em vez de um corte claro para o que conta como uma solução.

[355] Um defensor da abordagem de oráculo poderia insistir que ao menos existe a possibilidade de que o usuário percebesse a falha na solução apresentada — reconhecendo que ela não corresponderia diretamente às intenções do usuário, mesmo satisfazendo critérios de sucesso formalmente especificados. A probabilidade de se perceber o erro nesse estágio depende de vários fatores, incluindo o quanto as respostas do oráculo são humanamente compreensíveis e o quão caridoso ele é na seleção de quais características do resultado potencial são reveladas ao usuário. Por outro lado, em vez de confiar no próprio oráculo para fornecer essas funcionalidades, pode-se tentar construir uma ferramenta separada para fazer isso, uma ferramenta que possa inspecionar os pronunciamentos do oráculo e nos mostrar de forma útil o que aconteceria se atuássemos neles. Mas fazer isso com toda a generalidade exigiria outro oráculo superinteligente em cujas divinações teríamos que confiar; assim, o problema da confiabilidade não teria sido resolvido, apenas deslocado. Poder-se-ia procurar obter um aumento de segurança através do uso de múltiplos oráculos para realizar uma revisão pelos pares, mas isso não nos protegeria nos casos em que todos os oráculos falham da mesma maneira — o que pode acontecer se, por exemplo, todos tiverem recebido a mesma especificação formal do que conta como uma solução satisfatória.

[356] Com um poder computacional suficientemente grande — finito, porém fisicamente implausível —, *seria* provavelmente possível alcançar a superinteligência geral com os algoritmos disponíveis atualmente. (Cf., por exemplo, o sistema AIXI*tl;* Hutter [2001].) Mas mesmo a continuidade da lei de Moore por mais cem anos não seria suficiente para que obtivéssemos os níveis de poder computacional que poderiam tornar isso possível.

[357] Bird e Layzell (2002) e Thompson (1997); ver também Yaeger (1994, p. 13-14).

[358] Williams (1966).

[359] Leigh (2010).

[360] Esse exemplo foi retirado de Yudkowsky (2011).

[361] Wade (1976). Experimentos computacionais também foram conduzidos com uma evolução simulada cujo objetivo era reproduzir certos aspectos da evolução biológica — novamente, em certos casos, os resultados foram surpreendentes (ver, por exemplo, Yaeger [1994]).

[362] Esse tipo de cenário não é necessariamente o mais provável ou o mais desejado, mas é o mais fácil de analisar com as ferramentas oferecidas pela teoria econômica padrão, sendo, desse modo, um ponto de partida conveniente para nossa discussão.

[363] American Horse Council (2005). Ver também Salem e Rowan (2001).

[364] Acemoglu (2003); Mankiw (2009); Zuleta (2008).

[365] Fredriksen (2012, p. 8); Salverda et al. (2009, p. 133).

[366] Também é essencial que ao menos parte do capital seja investido em ativos cujo valor aumente de acordo com a tendência geral do mercado. Um portfólio diversificado de ativos, como ações em fundos indexados, aumentaria as chances de algum tipo de ganho.

[367] Muitos dos sistemas europeus de segurança social *não são capitalizados*, o que significa que os benefícios são pagos com os impostos e as contribuições dos trabalhadores atualmente ativos, e não com reservas de poupança. Tais esquemas não iriam automaticamente satisfazer o requisito — em casos repentinos de desemprego em massa, as receitas com as quais os benefícios fossem pagos poderiam secar. Entretanto, os governos poderiam optar por compensar esse déficit usando outras fontes.

[368] American Horse Council (2005).

[369] Fornecer uma pensão anual de 90 mil dólares a 7 bilhões de pessoas custaria 630 trilhões de dólares por ano, o que corresponde a dez vezes o valor atual do produto interno bruto mundial. Nos últimos cem anos, o produto interno bruto mundial cresceu aproximadamente dezenove vezes, partindo de 2 trilhões de dólares em 1900 para 37 trilhões de dólares em 2000 (em dólares internacionais de 1990), de acordo com Maddison (2007). Ou seja, se a taxa de crescimento que vimos nos últimos cem anos continuar pelos próximos duzentos anos, e caso a população permaneça constante, prover cada indivíduo com uma pensão de 90 mil dólares custaria aproximadamente 3% do produto interno bruto mundial. Uma explosão de inteligência poderia fazer com que esse crescimento econômico ocorresse num período de tempo muito mais curto. Ver também Hanson (1998a, 1998b, 2008).

[370] E talvez até 1 milhão de vezes mais nos últimos 70 mil anos, caso tenha havido um severo estreitamento da população nessa época, como tem sido especulado. Ver Kremer (1993) e Huff et al. (2010) para mais informações.

[371] Cochran e Harpending (2009). Ver também Clark (2007) e, para uma crítica, Allen (2008).

[372] Kremer (1993).

[373] Basten et al. (2013). Cenários nos quais há um crescimento contínuo também são possíveis. De forma geral, a incerteza a respeito de tais projeções aumenta consideravelmente se pensarmos além de uma ou duas gerações no futuro.

[374] Globalmente, em 2003, a taxa de fertilidade necessária para manter a população estável era de 2,33 crianças por mulher. Esse número se deve ao fato de que são necessárias duas crianças por mulher para a substituição dos pais e mais “um terço de uma criança” para compensar: (1) a maior probabilidade de nascer um menino, e (2) a morte precoce antes de atingir a fase fértil. Para países desenvolvidos, o número é menor — em torno de 2,1 — em virtude das menores taxas de mortalidade. (Ver Espenshade et al. [2003, Introdução, tabela 1, p. 580].) Não fossem os processos migratórios, a população na maioria dos países desenvolvidos diminuiria. Alguns exemplos importantes de países com taxas de fertilidade abaixo do nível necessário para substituição são: Singapura com 0,79 (a menor no mundo), Japão com 1,39, República Popular da China com 1,55, União Europeia com 1,58, Rússia com 1,61, Brasil com 1,81, Irã com 1,86, Vietnã com 1,87 e Reino Unido com 1,90. Até mesmo a população dos Estados Unidos provavelmente diminuiria ligeiramente com uma taxa de fertilidade de 2,05. (Ver CIA [2013]).

[375] A plenitude do tempo pode ocorrer daqui a muitos bilhões de anos.

[376] Carl Shulman destaca que, se humanos biológicos pretendem viver a totalidade de seu tempo de vida natural em uma economia digital, eles precisariam presumir não apenas que a ordem política nessa esfera digital protegeria os interesses humanos, mas que permaneceria assim por um longo período de tempo (Shulman, 2012). Por exemplo, se eventos na esfera digital transcorressem mil vezes mais rápido do que fora dela, então um humano biológico teria que confiar que a ordem política digital se manteria inalterada por 50 mil anos, sem qualquer mudança ou agitação interna. Entretanto, se a política digital tivesse qualquer semelhança

com a nossa, haveria um grande número de revoluções, guerras e levantes catastróficos durante esse intervalo de milênios, o que provavelmente seria um grande inconveniente para os humanos biológicos do lado de fora do mundo digital. Mesmo um risco de apenas 0,01% ao ano de uma guerra termonuclear ou catástrofe semelhante já significaria quase a certeza da perda total das vidas humanas que estivessem vivendo em câmera lenta no tempo sideral. Para superar esse problema, uma ordem mais estável no mundo digital seria necessária: talvez um *singleton* que gradualmente melhorasse sua própria estabilidade.

[377] É possível imaginar que, mesmo se as máquinas fossem muito mais eficientes que os humanos, ainda assim haveria algum nível de salário no qual seria lucrativo empregar trabalhadores humanos, digamos que a um centavo a hora. Se essa fosse a única fonte de renda para os humanos, nossa espécie se extinguiria, pois humanos não conseguem sobreviver com um centavo a hora. Mas humanos também recebem rendimentos de capital. Agora, se supormos que haveria crescimento populacional até que a renda total chegasse ao nível de subsistência, podemos imaginar que isso acarretaria uma situação na qual os humanos teriam que trabalhar muito. Por exemplo, suponhamos que a renda de subsistência seja de um dólar por dia. Nesse caso, ao que tudo indica, a população cresceria até que o capital per capita fornecesse apenas noventa centavos por dia de renda, a qual as pessoas teriam que complementar com dez horas de trabalho árduo para suprir os dez centavos restantes. Entretanto, o cenário não precisa ser esse, uma vez que o nível de subsistência depende da quantidade de trabalho realizado: humanos que trabalham arduamente queimam mais calorias. Suponhamos que cada hora trabalhada aumente o gasto com comida em dois centavos. Nesse caso, teríamos um modelo no qual os humanos permanecem ociosos em equilíbrio.

[378] Podemos pensar que um eleitorado tão enfraquecido como esse não seria capaz de votar ou defender suas posições. Mas aqueles que se encontrassem dentro desses “casulos” poderiam fornecer uma procuração para que IAs fiduciárias representassem seus interesses políticos. (Essa parte da discussão nesta seção está baseada na premissa de que os direitos de propriedade seriam respeitados.)

[379] Não é claro qual seria o melhor termo a ser usado aqui. “Matar” pode sugerir mais brutalidade do que implicaria um cenário real. Já “finalizar” pode soar como um eufemismo. Uma complicação é que existem, potencialmente, dois eventos distintos: o de parar de rodar ativamente um processo e o de apagar as informações contidas em um modelo. Ambos normalmente estão presentes na morte de um ser humano, mas, para emulações, eles podem se dar em separado. O fato de um programa parar temporariamente de rodar pode não ter grandes consequências, sendo apenas equivalente a um período de sono de um humano, mas finalizá-lo permanentemente poderia se equivaler a um coma profundo. Surgem ainda mais complicações ligadas ao fato de que emulações podem ser copiadas e rodadas a velocidades diferentes: possibilidades que não têm equivalente direto na experiência humana. (Cf. Bostrom [2006b]; Bostrom e Yudkowsky [2015].)

[380] Haverá uma relação inversa entre o poder computacional paralelo total e a velocidade computacional, uma vez que as maiores velocidades computacionais somente serão atingidas ao custo de uma redução na eficiência energética. Isso será especialmente verdadeiro depois que entrarmos na era da computação reversível.

[381] Um teste poderia ser feito induzindo a emulação a uma tentação. Testando repetidamente como uma emulação que parta de um estado predefinido reagiria a várias sequências de estímulos, poderíamos saber se essa emulação teria um alto grau de confiabilidade. Mas quanto mais for permitido um posterior desenvolvimento desse estado mental para além de seu estágio inicial validado, menor seria a certeza que teríamos de que ela se manteria confiável. (Particularmente, como uma emulação astuta poderia se dar conta de que ela se encontra em uma simulação, precisaríamos ser cautelosos ao extrapolar seu comportamento para situações onde sua hipótese de simulação teria menos peso na sua tomada de decisões.)

[382] Algumas emulações poderiam ter uma identificação com seu clã — ou seja, com todas as suas cópias e variações derivadas do mesmo modelo — em vez de se identificar com uma instanciação em particular. Uma emulação assim poderia não considerar o seu fim como uma morte, se ela soubesse que outros membros do clã sobreviveriam. As emulações poderiam se dar conta de que elas seriam revertidas a um estado particular predefinido ao fim do expediente e perder a memória adquirida durante esse dia, mas viveriam isso como alguém que bebe muito em uma festa e acorda sem se lembrar do que fez na noite anterior, ou seja, como uma amnésia retrógrada, e não como a morte.

[383] Uma avaliação ética também pode levar em consideração muitos outros fatores. Mesmo que todos os trabalhadores estivessem constantemente satisfeitos com sua condição, o resultado ainda assim poderia ser moralmente questionável a partir de outros pontos de vista — a definição de quais outros pontos de vista, entretanto, é uma questão disputada entre teorias morais rivais. Mas qualquer avaliação plausível consideraria o bem-estar subjetivo como um fator importante. Ver também Bostrom e Yudkowsky (2015).

[384] World Values Survey (2008).

[385] Helliwell e Sachs (2012).

[386] Cf. Bostrom (2004). Ver também Chislenko (1996) e Moravec (1988).

[387] É difícil dizer se as estruturas de processamento de informação que surgiriam nesse tipo de cenário seriam conscientes (no sentido de possuírem *qualia*, experiência fenomenal). Essa é uma questão difícil em parte por nossa ignorância empírica a respeito de quais entidades cognitivas surgiriam e também em virtude da nossa ignorância filosófica em relação a quais tipos de estrutura possuem consciência. Poderíamos tentar reformular a questão da seguinte maneira: em vez de perguntar se as entidades futuras teriam ou não consciência, poderíamos questionar se elas possuiriam ou não status moral; ou se elas apresentariam certas características que nos fariam ter algum tipo de preferência em relação ao seu "bem-estar". Mas essas perguntas podem ser tão difíceis quanto questões a respeito da consciência — na verdade, elas poderiam exigir uma resposta para a questão da consciência, pois o status moral ou nossas preferências dependem da capacidade da entidade em questão de experienciar subjetivamente sua condição.

[388] Para um argumento a favor de que tanto a geologia como a história humana manifestam essa tendência em direção a uma maior complexidade, ver Wright (2001). Para um argumento oposto (criticado no capítulo 9 da obra de Wright), ver Gould (1990). Ver também Pinker (2011), segundo o qual estamos testemunhando uma forte tendência de longo prazo em direção a uma diminuição da violência e da brutalidade.

[389] Mais sobre a teoria de seleção por observação, ver Bostrom (2002a).

[390] Bostrom (2008a). Seria necessária uma análise muito mais cuidadosa dos detalhes de nossa história evolutiva para eliminar o efeito de seleção. Ver, por exemplo, Carter (1983, 1993); Hanson (1998d); Ćirković et al. (2010).

[391] Kansa (2003).

[392] Ver, por exemplo, Zahavi e Zahavi (1997).

[393] Ver Miller (2000).

[394] Kansa (2003). Para uma visão mais provocativa, ver também Frank (1999).

[395] Não é evidente qual seria a melhor maneira de medir o grau de integração política global. Enquanto no passado uma tribo de caçadores-coletores integrava cem caçadores individuais em uma entidade que tomava decisões em conjunto, atualmente as maiores entidades políticas contam com mais de 1 bilhão de indivíduos. Essa é uma diferença que atinge sete ordens de magnitude. Apenas uma ordem de magnitude a mais faria com que toda a população do planeta estivesse contida em uma única entidade política. Entretanto, no período em que a tribo representava a maior escala de integração existente, a população mundial era muito menor. Uma tribo pode ter contido cerca de um milésimo da quantidade total de indivíduos do período. Visto dessa forma, o aumento na escala de integração política seria de apenas duas ordens de magnitude. Olhar para a fração da população mundial que se encontra politicamente integrada, em vez de considerar seu número absoluto, parece apropriado dentro do presente contexto (particularmente porque a transição para a inteligência de máquina pode fazer com que haja uma explosão populacional de emulações ou outras mentes digitais). Mas também houve o desenvolvimento de instituições globais e redes de colaboração para além das estruturas formais dos Estados, as quais também devem ser levadas em consideração.

[396] Uma das razões para supor que a primeira revolução de inteligência de máquina será veloz — a possível existência de um excesso de hardware — não se aplica aqui. Entretanto, podem existir outras fontes de ganho rápido, tal como um progresso extraordinário em relação ao desenvolvimento de software associado à transição de emulações para inteligências de máquina puramente sintéticas.

[397] Shulman (2010b).

[398] Como os prós e contras se equilibrariam, pode depender do tipo de trabalho que o superorganismo estivesse tentando fazer e o quão capaz em termos gerais seria a emulação mais capaz usada como modelo. Em parte, diferentes tipos de seres humanos são necessários em grandes organizações atualmente porque humanos muito talentosos em diversas áreas do conhecimento são raros.

[399] É claro que fazer múltiplas cópias de um mesmo agente de software é uma tarefa fácil. Mas note que, geralmente, apenas criar cópias não é suficiente para assegurar que elas tenham o mesmo objetivo final. Para que dois agentes tenham os mesmos objetivos finais (no sentido relevante de "mesmo"), esses objetivos precisam ter os mesmos elementos indexicais. Se João é egoísta, uma cópia do João será igualmente egoísta. Contudo, seus objetivos não coincidem: João se preocupa apenas consigo mesmo, enquanto a cópia do João se preocupa apenas com a cópia do João.

[400] Shulman (2010b, p. 6).

[401] Isso pode ser mais viável para humanos biológicos e emulações completas do cérebro do que para inteligências artificiais arbitrárias, que podem ter sido construídas com compartimentos secretos ou dinâmicas funcionais muito difíceis de descobrir. Por outro lado, *IAs construídas especificamente para ser transparentes* deveriam permitir uma inspeção e uma verificação mais detalhada do que aquela que é possível fazer em arquiteturas semelhantes ao cérebro. Pressões sociais podem encorajar as IAs a exporem seu código-fonte e a

modificarem a si mesmas de modo a tornar sua arquitetura transparente — especialmente se isso for uma pré-condição para que elas se tornem confiáveis e, assim, lhes seja dada a oportunidade de participar de transações benéficas. Cf. Hall (2007).

[402] Outras questões que parecem ser relativamente menos importantes, especialmente em situações nas quais o que está em jogo é crucial (como no caso de falhas de coordenação global), incluem o custo de procura por políticas que possam ser de interesse mútuo e a possibilidade de que alguns agentes apresentem uma preferência básica por "autonomia", que seria mitigada pela adesão a tratados globais abrangentes que possuíssem mecanismos próprios de monitoramento e cumprimento de regras.

[403] É possível que uma IA seja capaz de fazer isso modificando a si mesma de forma apropriada e então fornecendo a seus observadores um acesso estritamente de leitura a seu código-fonte. Uma inteligência de máquina com uma arquitetura mais opaca (tal como uma emulação) poderia, talvez, conseguir isso aplicando publicamente em si mesma algum método de seleção de motivação. Alternativamente, uma agência externa de coerção, como um superorganismo que funcione como uma força policial, poderia, talvez, ser usado não apenas para impor a implementação de um tratado acordado entre partes diferentes, mas também internamente, fazendo com que uma única parte se comprometa a um curso particular de ação.

[404] A seleção evolutiva pode ter favorecido pessoas que ignoram ameaças e mesmo aquelas visivelmente tão inflexíveis que prefeririam lutar até a morte do que sofrer a menor das perdas. Tal temperamento pode fornecer ao seu possuidor alguns benefícios valiosos de sinalização. (Qualquer recompensa instrumental de se possuir esse temperamento não precisa, é claro, ter nenhum papel na motivação consciente do agente: ele poderia valorizar a honra e a justiça como fins em si mesmas.)

[405] Entretanto, um veredito definitivo a respeito dessas questões deve esperar por novas análises. Existem diversas outras complicações em potencial que não poderemos explorar aqui.

[406] Diversas complicações e modulações podem ser introduzidas a partir dessa ideia básica. Discutimos uma de suas variações no capítulo 8 — a ideia de um agente que se satisfaça em oposição a um agente de maximização — e no próximo capítulo falaremos brevemente sobre a questão das teorias alternativas de decisão. Entretanto, tais questões não são essenciais para o teor desta subseção, de modo que tentaremos manter tudo o mais simples possível, voltando nossa atenção para o caso específico de um agente maximizador de utilidades esperadas.

[407] Considerando que a IA deva ter uma função de utilidade não trivial. Seria muito fácil construir um agente que sempre escolha uma ação que maximize a utilidade esperada se a sua função de utilidade for, por exemplo, a função constante U(w) = 0. Qualquer ação maximizaria igualmente bem a utilidade esperada em relação a essa função de utilidade.

[408] E também porque já nos esquecemos da confusão "florescente e zumbidora" que isso representava nos estágios iniciais de nossas infâncias, um período em que ainda não conseguíamos enxergar muito bem porque nossos cérebros ainda não tinham aprendido como interpretar dados de entrada visuais.

[409] Ver também Yudkowsky (2011) e a resenha na seção 5 de Muehlhauser e Helm (2012).

[410] É talvez quase concebível que avanços em engenharia de software possam, em algum momento, superar essas dificuldades. Utilizando-se de ferramentas modernas, um único programador é capaz de produzir um software que seria incomparável a softwares feitos por grandes equipes de programadores que fossem forçados a programar diretamente em código de máquina. Os atuais programadores de IA ganharam expressividade em virtude da ampla variedade de técnicas de aprendizagem de máquina e bibliotecas de matemática científica disponíveis, possibilitando o desenvolvimento, por exemplo, de um novo sistema de contagem de rostos via webcam através da junção de várias dessas bibliotecas, algo que eles não conseguiriam desenvolver por conta própria partindo do zero. A acumulação de softwares reutilizáveis produzidos por especialistas, mas que podem ser utilizados por não especialistas, fornecerá uma vantagem expressiva a futuros programadores. Por exemplo, alguém que decida no futuro desenvolver um programa de robótica poderá ter acesso rápido a bibliotecas padrão de reconhecimento facial, coleções de objetos comuns em escritórios, bibliotecas de trajetórias especializadas e muitas outras funcionalidades atualmente indisponíveis.

[411] Dawkins (1995, p. 132). A afirmação aqui não é necessariamente que a quantidade de sofrimento no mundo natural supera a quantidade de bem-estar.

[412] Os tamanhos de população necessários podem ser muito maiores ou muito menores do que aqueles que existiram na nossa própria ascendência. Ver Shulman e Bostrom (2012).

[413] Se fosse fácil alcançar um resultado semelhante sem prejudicar um grande número de vidas inocentes, isso seria moralmente aconselhável. Se, contudo, seres digitais fossem criados e sofressem injustamente, talvez esse sofrimento pudesse ser compensado salvando-os em arquivos para, mais tarde (quando o futuro da humanidade estiver assegurado), rodá-los novamente sob condições mais favoráveis. Poderíamos, de alguma

maneira, comparar tais restituições com concepções religiosas de vida após a morte, dentro do contexto de tentativas teológicas de explicar a questão do mal.

[414] Uma das principais autoridades nesse assunto, Richard Sutton, define a aprendizagem por reforço não como um método de aprendizagem, mas como um problema de aprendizagem: qualquer método que seja adequado para resolver o problema em questão é considerado um método de aprendizagem por reforço (Sutton e Barto, 1998, p. 4). Em contraste, a presente discussão se atém aos métodos nos quais o agente pode ser concebido como dotado do objetivo final de maximizar (alguma noção de) recompensa cumulativa. Uma vez que um agente que possua um objetivo final muito distinto poderia ter a capacidade de imitar um agente que busque recompensas em diversas situações diferentes, sendo, dessa forma, capaz de resolver problemas de aprendizagem por reforço, é possível que existam outros métodos que contariam como "métodos de aprendizagem por reforço", de acordo com a definição de Sutton, que não resultariam na síndrome de *wireheading*. As observações feitas no texto, entretanto, se aplicam à maioria dos métodos de fato utilizados por pesquisadores que estudam a aprendizagem por reforço.

[415] Mesmo se, de algum modo, um mecanismo semelhante ao humano pudesse ser programado para funcionar em uma máquina com intelecto semelhante ao humano, os objetivos finais adquiridos por esse intelecto podem não se parecer com aqueles de um humano normal, a não ser que o ambiente onde ele tivesse se desenvolvido em sua infância fosse muito próximo àquele de uma criança comum: algo que seria complicado de replicar. E mesmo em um ambiente semelhante ao humano um resultado satisfatório não seria garantido, uma vez que qualquer diferença, por mais sutil, poderia resultar em reações muito diferentes a um evento vivido. Pode acontecer, entretanto, que seja possível criar um mecanismo de acréscimo de valores mais confiável para mentes semelhantes às humanas no futuro (talvez com o uso de novas drogas e implantes cerebrais, ou mesmo seu equivalente digital).

[416] É possível se perguntar por que nós, humanos, não estamos tentando desativar o mecanismo que nos conduz à aquisição de novos valores finais. Diversos fatores podem estar em jogo aqui. Em primeiro lugar, o sistema de motivação humano pode ser descrito, de maneira muito simplória, como um algoritmo que calcula friamente uma maximização de utilidade. Em segundo lugar, talvez não tenhamos nenhuma maneira conveniente de alterar a forma como adquirimos valores. Em terceiro lugar, talvez tenhamos razões instrumentais (que surjam, por exemplo, da necessidade de sinalização social) para, em alguns momentos, adquirir novos valores finais — valores instrumentais podem não ser tão úteis se nossas mentes forem parcialmente transparentes para outras pessoas, ou se a complexidade cognitiva de fingir possuir valores diferentes daqueles que realmente temos for muito alta. Em quarto lugar, existem casos em que nós resistimos ativamente a tendências que produzem mudanças em nossos valores finais, por exemplo, quando buscamos resistir à influência nociva de más companhias. E, por último, existe a interessante possibilidade de atribuirmos algum valor final em nos manter como o tipo de agente que pode adquirir novos valores finais de maneiras humanas consideradas normais.

[417] Poderíamos tentar também desenvolver um sistema de motivação no qual a IA fosse indiferente a tal substituição; ver Armstrong (2010).

[418] Nos basearemos aqui nas elucidações feitas por Daniel Dewey (2011). Outras ideias que contribuíram para essa formulação foram desenvolvidas por Marcus Hutter (2005) e Shane Legg (2008), Eliezer Yudkowsky (2001), Nick Hay (2005), Moshe Looks e Peter de Blanc.

[419] Para evitar complicações desnecessárias, nossa análise ficará limitada a agentes determinísticos que não descontam recompensas futuras.

[420] Matematicamente, o comportamento de um agente pode ser formalizado como uma *função agente*, que mapeia cada histórico possível de interações a uma ação. Com exceção de agentes extremamente simples, é impraticável representar uma função agente explicitamente em forma tabular. Em vez disso, seria fornecida ao agente alguma maneira de calcular qual ação ele deveria executar. Como existem muitas maneiras de calcular a mesma função agente, isso levaria a uma individuação mais minuciosa de um agente na forma de um *programa agente*. Um programa agente é um programa ou algoritmo específico que calcula uma ação para qualquer histórico de interações dado. Apesar de ser, em geral, matematicamente conveniente e útil pensar em um programa agente que interaja com algum tipo de ambiente formalmente especificado, é importante lembrar que isso é uma idealização. Agentes reais são fisicamente instanciados. Isso significa não apenas que o agente interage fisicamente com o meio ambiente por meio de sensores e atuadores, mas também que o "cérebro" ou o controlador do agente é, *ele mesmo, parte da realidade física*. Desse modo, suas operações podem, em princípio, sofrer interferência física externa (não apenas através de percepções recebidas por seus sensores). Sendo assim, em algum momento, torna-se necessário ver o agente como uma *implementação de agente*. Uma implementação de agente é uma estrutura física que, na ausência de interferência do ambiente ao redor, implementa uma função de agente. (Essa definição segue Dewey [2011].)

[421] Dewey propôs a seguinte noção de otimalidade para um agente de inserção de valores:

$$y_k = \arg\max_{y_k} \sum_{x_k yx_{k+1:m}} P_1(yx_{\leq m} | yx_k y_k) \sum_U U(yx_{\leq m}) P_2(U | yx_{\leq m}).$$

Aqui, P_1 e P_2 são duas funções de probabilidade. O segundo somatório varia sobre alguma classe adequada de funções de utilidade dos históricos de interação possíveis. Na versão apresentada no corpo do texto, tornamos explícitas algumas dependências, bem como nos beneficiamos da notação simplificada de mundos possíveis.

[422] Devemos notar que o conjunto de funções de utilidade U deve permitir que as utilidades possam ser comparadas e ponderadas. Isso é, de modo geral, problemático, e a representação de diferentes teorias morais do bem em termos de funções de utilidade cardinais não é sempre óbvia. (Ver, por exemplo, MacAskill, [2010].)

[423] Ou, de modo mais geral, como V pode não implicar diretamente que a proposição V (*U*) seja verdadeira em *w*, para qualquer par formado por um mundo possível e uma função de utilidade (*w, U*), é preciso dar à IA uma representação adequada da distribuição de probabilidade condicional *P* (V (*U*)|*w).*

[424] Considere em primeiro lugar Y, que corresponde à classe de ações possíveis a um agente. Um dos problemas aqui consiste em saber o que exatamente contaria como uma ação: apenas comandos motores básicos (por exemplo, “envie um pulso elétrico através do canal de saída #00101100”) ou ações mais complexas (por exemplo, “mantenha a câmera centralizada na face”)? Como o que estamos tentando desenvolver é uma noção de otimalidade, e não um plano de implementação prático, podemos escolher aqui os comandos motores básicos (e já que o conjunto de possíveis comandos motores pode mudar ao longo do tempo, talvez seja necessário indexar Y à variável de tempo). Entretanto, para seguirmos em direção a uma implementação, será presumivelmente necessário que se introduza algum tipo de processo de planejamento hierárquico e talvez precisemos considerar como aplicar essa fórmula a alguma classe de ações mais complexas. Outra questão é como analisar ações internas (tais como a escrita de sequências de dados na memória). Uma vez que ações internas podem ter consequências importantes, seria ideal que Y incluísse também essas ações internas básicas, e não apenas comandos motores. Mas há limites em relação a quão longe poderíamos ir nessa direção: o cálculo da utilidade esperada de qualquer ação em Y requer múltiplas operações computacionais, e se cada uma dessas operações também tiver que ser considerada como uma ação interna em Y que precisará ser avaliada pela IA-IV estaríamos, então, diante de uma regressão infinita que tornaria impossível sequer começar. Para evitar essa regressão infinita, poderíamos restringir qualquer tentativa explícita de estimar a utilidade esperada a um número limitado de possibilidades significativas de ação. O sistema precisaria, então, de uma espécie de processo heurístico que identificasse algumas possibilidades significativas de ação para fazer maiores considerações. (Posteriormente, o sistema também poderia ser capaz de tomar decisões explícitas em relação a algumas possíveis ações que modificassem esse processo heurístico, ações que poderiam ter sido identificadas para receber atenção explícita dentro desse mesmo processo; desse modo, no longo prazo, o sistema poderia se tornar cada vez mais efetivo em aproximar aquilo que a IA-IV consideraria como ideal.)

Considere agora W, que é uma classe de mundos possíveis. Uma dificuldade aqui consiste em especificar W de modo que ele seja suficientemente inclusivo. Se falhássemos em incluir algum w relevante em W, a IA se tornaria incapaz de representar uma situação que realmente ocorresse, o que a faria tomar decisões ruins. Suponha, por exemplo, que usássemos alguma teoria ontológica para determinar os elementos de W. Por exemplo, se incluíssemos em W todos os mundos possíveis que consistissem de um certo tipo de superfície espaço-tempo, compostos por partículas elementares que são encontradas na teoria padrão da física de partículas. Isso poderia distorcer a epistemologia da IA caso a teoria padrão estivesse incompleta ou incorreta. Poderíamos tentar utilizar uma classe maior para W, o que possibilitaria cobrir mais possibilidades. Mas mesmo que fosse possível assegurar a inclusão de todos os universos físicos, ainda assim poderíamos nos preocupar com a eventualidade de que alguma outra possibilidade tenha sido deixada de fora. Por exemplo, o que fazer com a possibilidade de existirem mundos possíveis dualísticos onde os fatos a respeito da consciência não advêm de fatos a respeito da física? E o que fazer em relação aos fatos indexicais? E quanto aos fatos normativos? E os fatos de matemática avançada? E em relação a outros tipos de fato que nós, humanos falíveis, pudéssemos ter ignorado, mas que seriam importantes para que tudo corresse da melhor maneira possível? Algumas pessoas têm grande convicção de que alguma teoria ontológica específica estaria correta. (Entre aqueles que escrevem a respeito do futuro da IA, a crença em uma ontologia materialista, segundo a qual fatores mentais advêm de fatores físicos, é, em geral, tida como algo certo.) Contudo, uma rápida reflexão sobre a história das ideias deveria nos ajudar a entender que existe uma possibilidade significativa de que nossa ontologia favorita esteja errada. Caso os cientistas do século XIX tivessem tentado elaborar uma definição de W inspirada na física, eles provavelmente teriam sido negligentes em incluir a possibilidade da existência de um espaço-tempo não euclidiano ou de uma interpretação da teoria quântica segundo Everett (“muitos mundos”),

ou de um multiverso cosmológico, ou da hipótese de simulação — possibilidades que atualmente parecem ter uma probabilidade substancial de serem verificadas no mundo real. É plausível que existam outras possibilidades que nossa geração atual também desconheça. (Por outro lado, se W for muito grande, surgem dificuldades técnicas relativas à necessidade de definir medidas para conjuntos transfinitos.) O ideal seria que fôssemos, de alguma forma, capazes de organizar tudo de modo que a IA pudesse usar algum tipo de ontologia aberta, uma que a própria IA, subsequentemente, fosse capaz de expandir, usando para tanto os mesmos princípios que nós, humanos, aplicaríamos ao decidir reconhecer algum novo tipo de possibilidade metafísica.

Considere P ($w|Ey$). Especificar essa probabilidade condicional não é estritamente parte do problema de inserção de valores. Para que a IA possa ser inteligente, ela precisa ser capaz, de antemão, de calcular probabilidades razoavelmente precisas a respeito de muitas possibilidades factuais relevantes. Um sistema que seja incapaz de fazer isso não apresentará os tipos de perigo com os quais nos preocupamos aqui. No entanto, pode existir o risco de que a IA acabe tendo uma epistemologia que seja boa o suficiente para torná-la instrumentalmente efetiva, mas não boa o suficiente para capacitá-la a pensar de forma correta a respeito de algumas possibilidades que tenham grande importância normativa. (O problema de se especificar P ($w|Ey$) estaria, assim, relacionado à questão da especificação de W.) Especificar P ($w|Ey$) também requer que confrontemos outras questões, por exemplo: como representar incerteza sobre impossibilidades lógicas.

As questões mencionadas acima — a respeito de como definir uma classe de ações possíveis, uma classe de mundos possíveis e uma distribuição de probabilidade conectando evidências a classes de mundos possíveis — são bastante genéricas: questões similares surgem para uma ampla gama de agentes especificados formalmente. Resta agora examinar um conjunto de questões mais peculiares à abordagem de aprendizagem de valores, especificamente: como definir U, V (U) e P (V (U) | w).

U é uma classe de funções de utilidade. Existe uma ligação entre U e W, na medida em que cada função de utilidade U (w) em U deveria, idealmente, atribuir utilidades a cada mundo possível w em W. Mas U também precisa ser amplo, no sentido de que deveria conter uma quantidade suficientemente grande e diversa de funções de utilidade para que possamos justificar nossa confiança de que ao menos uma dessas funções de utilidade seja efetiva em representar os valores pretendidos.

Devemos escrever P (V (U) | w), em vez de simplesmente P (U |w), para enfatizar o fato de que probabilidades são atribuídas a proposições. Uma função de utilidade não é em si uma proposição, mas podemos transformar uma função de utilidade em uma proposição se fizermos algumas afirmações a seu respeito. Por exemplo, podemos afirmar que uma função de utilidade específica U (.) descreve as preferências de uma pessoa em particular, ou que representa as prescrições derivadas de alguma teoria ética, ou que é a função de utilidade que o diretor teria desejado implementar caso tivesse pensado meticulosamente. O "critério de valor" V (.) poderia ser interpretado, então, como uma função que recebe como entrada uma função de utilidade U e devolve como saída uma proposição indicando se U satisfaz o critério V. Uma vez que tenhamos definido uma proposição V (U), poderíamos, quem sabe, obter a probabilidade condicional P (V (U)|w) a partir de qualquer fonte que usássemos para obter as outras distribuições de probabilidade na IA. (Se tivéssemos certeza de que todos os fatos normativamente relevantes estariam sendo levados em consideração na individualização dos mundos possíveis W, então P (V (U)|w) deveria ser igual a zero (0) ou a um (1) em qualquer um dos mundos possíveis.) A questão que permanece é a de como definir V. Isso será discutido mais adiante no texto.

[425] Esses não são os únicos desafios à abordagem de aprendizado de valores. Outra questão, por exemplo, seria como fazer com que a IA possuísse crenças iniciais suficientemente sensatas — pelo menos até o momento em que ela se tornasse forte o suficiente para subverter as tentativas de correção por parte dos programadores.

[426] Yudkowsky (2001).

[427] Cf. capítulos 9 e 11.

[428] Por exemplo, o MDMA pode, temporariamente, aumentar a empatia; a oxitocina é capaz de aumentar temporariamente a sensação de confiança (Vollenweider et al., [1998]; Bartz et al., [2011]). No entanto, seus efeitos parecem variar bastante e dependem muito do contexto.

[429] O termo "ave-maria" [tradução do inglês de *Hail Mary*] foi retirado do futebol americano e faz referência a um passe muito longo feito em um momento de desespero, em geral quando o tempo de jogo está se esgotando, na esperança de que um parceiro de equipe possa apanhar a bola perto da linha de fundo e marcar um *touchdown*.

[430] A abordagem “ave-maria” é baseada na ideia de que uma superinteligência seria capaz de articular suas preferências com uma exatidão muito maior do que nós, humanos, conseguimos articular as nossas. Por exemplo, uma superinteligência poderia especificar suas preferências em forma de código. Desse modo, se nossa superinteligência estivesse representando outras superinteligências na forma de processos computacionais que estivessem percebendo seu ambiente, então nossa IA deveria ser capaz de compreender como uma superinteligência alienígena responderia a algum estímulo hipotético, tal como uma “janela” que surgisse em seu campo visual mostrando o código-fonte de nossa IA e solicitando que a IA alienígena especificasse suas instruções em algum formato conveniente predeterminado. Nossa IA poderia, então, fazer uma leitura dessas instruções imaginárias (a partir de seu próprio modelo desse cenário contrafactual no qual essas superinteligências alienígenas estariam representadas) e nós a teríamos construído de modo que ela estivesse motivada a seguir tais instruções.

[431] Uma alternativa seria criar um detector que procurasse (dentro do modelo de mundo de nossa IA) por (representações de) estruturas físicas criadas por uma civilização superinteligente. Desse modo, evitaríamos a necessidade de identificar as funções de preferência dessa superinteligência hipotética e daríamos à nossa IA um valor final que consistiria em tentar copiar quaisquer estruturas físicas que ela acreditasse que civilizações superinteligentes tenderiam a construir.

Entretanto, também existem desafios técnicos nessa versão. Por exemplo, uma vez que nossa IA, mesmo depois de ter alcançado a superinteligência, talvez não fosse capaz de saber com muita precisão quais estruturas físicas outras superinteligências construiriam, ela poderia ter que tentar fazer uma aproximação a respeito dessas estruturas. Para tanto, é provável que nossa IA necessitasse de uma métrica de similaridade por meio da qual pudesse julgar quão semelhante um artefato físico seria em relação a outro. Porém métricas de similaridade baseadas em medidas físicas brutas podem ser inadequadas — por exemplo, não haveriam nenhuma utilidade em julgar que um cérebro se parece mais com um queijo camembert do que com um computador rodando uma emulação.

Mais viável seria uma abordagem que procurasse por “sinais”: mensagens a respeito de funções de utilidade codificadas em algum tipo de formato simples e adequado. Poderíamos construir nossa IA para que ela seguisse quaisquer dessas mensagens sobre funções de utilidade que hipoteticamente poderiam existir no universo; e teríamos que ter esperança de que IAs extraterrestres amigáveis criariam uma variedade de “sinais” que elas acreditassem (com base em sua superinteligência) que poderiam ser procurados por IAs que seriam provavelmente construídas por civilizações simples como a nossa.

[432] Se *todas* as outras civilizações tentassem resolver o problema da inserção de valores através de uma abordagem “ave-maria”, nossa tentativa falharia. Alguém teria que resolver esse problema do jeito difícil.

[433] Christiano (2012).

[434] A IA que seria construída também não precisaria ser capaz de encontrar o modelo. Assim como nós, ela poderia refletir a respeito das implicações de tal definição implícita complexa (talvez observando seu ambiente e seguindo basicamente a mesma linha de raciocínio que nós, humanos, seguiríamos).

[435] Os subagentes melhorados poderiam ser mortos ou colocados em um estado de animação suspensa (pausados), ser revertidos a um estado anterior às melhorias ou destituídos e impedidos de receber qualquer tipo de melhoria adicional até que todo o sistema tivesse alcançado um estágio de maior maturidade e segurança, no qual esse primeiro subgrupo melhorado não representasse mais risco para o sistema como um todo.

[436] A questão pode ser ainda menos óbvia em uma sociedade futura de humanos biológicos, que tenha acesso a técnicas avançadas de vigilância ou técnicas biomédicas para manipulação psicológica, ou que seja rica o suficiente para ter condições de contratar um número extremamente alto de profissionais de segurança que vigiassem os cidadãos comuns (e também uns aos outros).

[437] Cf. Armstrong (2007) e Shulman (2010b).

[438] Uma questão que segue em aberto é: até que ponto um supervisor de nível n precisaria monitorar não apenas agentes do nível (n — 1), mas também agentes monitorados por eles, do nível (n — 2), para ter certeza de que os agentes de nível (n — 1) estivessem cumprindo efetivamente com suas funções. E para verificar se os agentes de nível (n — 1) estariam sendo bem-sucedidos no gerenciamento de agentes de nível (n — 2), não seria também necessário que esse agente de nível n monitorasse os agentes de nível (n — 3)?

[439] Essa abordagem transpõe a linha entre a seleção de motivação e o controle de capacidade. Tecnicamente, a parte do arranjo que consiste de seres humanos controlando um grupo de softwares responsáveis pela supervisão do sistema poderia ser vista como controle de capacidade, enquanto a parte que consiste em camadas de agentes de software controlando outras camadas de agentes dentro do sistema

poderia ser considerada como um processo de seleção de motivação (conquanto que esse seja um mecanismo capaz de moldar as tendências motivacionais do sistema).

[440] Na verdade, também seria válido considerar muitos outros, mas não podemos fazê-lo aqui. Por exemplo, quaisquer agentes que fossem encarregados de comandar tal hierarquia poderiam ser corrompidos ou degradados pelo seu poder.

[441] Para que essa garantia fosse de fato efetiva, ela teria que ser implementada de boa--fé. Isso descartaria certos tipos de manipulação das emoções e da capacidade de tomada de decisão das emulações, que de outro modo poderiam ser usadas para (por exemplo) criar um temor na emulação de que ela poderia ser eliminada ou para impedir que ela considerasse racionalmente todas as suas opções.

[442] Ver, por exemplo, Brinton (1965); Goldstone (1980, 2001). (O progresso nas ciências sociais em relação a essas questões poderia ser um belo presente aos déspotas do mundo, que poderiam usar previsões mais precisas de agitação social para otimizar suas estratégias de controle da população e eliminar insurgências de forma mais suave e com menos força letal, enquanto elas ainda se encontrassem em seus estágios iniciais.)

[443] Cf. Bostrom (2011a, 2009b).

[444] No caso de um sistema inteiramente artificial, talvez fosse possível obter algumas das vantagens de uma estrutura institucional sem a real necessidade de criar subagentes distintos. Um sistema poderia incorporar múltiplas perspectivas em seus processos decisórios sem dotar cada uma dessas perspectivas com a totalidade das faculdades cognitivas necessárias para a atuação independente. Entretanto, poderia ser complicado implementar completamente a funcionalidade descrita no texto “observar as consequências comportamentais de uma modificação proposta e revertê-la a uma versão anterior caso as consequências se mostrassem indesejáveis previamente” em um sistema que não seja composto por subagentes.

[445] Uma pesquisa recente realizada com filósofos profissionais apresentou porcentagens de entrevistados que “aceitaram ou tinham inclinação a aceitar” diversas posições. Em relação à ética normativa, os resultados foram: 25,9% deontologia; 23,6% consequencialismo; 18,2% ética da virtude. Na metaética, os resultados foram: 56,4% realismo moral; 27,7% antirrealismo moral. Em julgamento moral: 65,7% cognitivismo; 17% não cognitivismo (Bourget e Chalmers, 2009).

[446] Pinker (2011).

[447] Para uma discussão sobre esse assunto, ver Shulman et al. (2009).

[448] Moore (2011).

[449] Bostrom (2006b).

[450] Bostrom (2009b).

[451] Bostrom (2011a).

[452] Nós deveríamos, mais precisamente, optar pela opinião da superinteligência, exceto quando tivéssemos boas razões para supor que nossas crenças seriam mais acertadas em relação a determinados tópicos. Por exemplo, talvez pudéssemos saber mais do que a superinteligência a respeito do que estivéssemos pensando em um determinado momento, caso ela não fosse capaz de ler nossos cérebros. Todavia, nós poderíamos omitir essa qualificação se assumíssemos que a superinteligência teria acesso às nossas opiniões; então, seria possível deixar também a cargo da superinteligência a tarefa de julgar quando confiar em nossas opiniões. (Poderiam restar alguns casos especiais que envolveriam informações indexicais, as quais precisariam ser abordadas separadamente — por exemplo, fazer com que a superinteligência explicasse no que seria racional acreditar de acordo com nossas perspectivas.) Para uma introdução à crescente literatura filosófica a respeito de testemunhos e autoridade epistêmica, ver, por exemplo, Elga (2007).

[453] Yudkowsky (2004). Ver também Mijic (2010).

[454] David Lewis, por exemplo, propôs uma teoria disposicional de valor, que sustenta, grosso modo, que X é um valor para A se, e somente se, A desejasse querer X se A fosse perfeitamente racional e idealmente familiarizada com X (Smith et al., 1989). Ideias semelhantes foram apresentadas anteriormente; ver, por exemplo, Sen e Williams (1982), Railton (1986), e Sidgwick e Jones (2010). Seguindo uma linha parecida, uma explicação comum de justificação filosófica, o método do equilíbrio reflexivo propõe um processo de ajuste mútuo iterativo entre as nossas intuições a respeito de casos específicos, as regras gerais que achamos que governam esses casos e os princípios segundo os quais nós pensamos que esses elementos devem ser revistos, como forma de alcançar um sistema mais coerente; ver, por exemplo, Rawls (1971) e Goodman (1954).

[455] A intenção aqui é que, presumivelmente, quando a IA atuasse para prevenir esse tipo de desastre, ela deveria fazê-lo da forma mais suave possível, ou seja, de uma tal maneira que evitasse o desastre, mas sem exercer demasiada influência sobre como seria o futuro da humanidade em outros aspectos.

[456] Yudkowsky (2004).

[457] Rebecca Roache, comunicação pessoal.

[458] Os três princípios são: “Defender os humanos, o futuro da humanidade e a natureza humanista” (humanista aqui se refere àquilo que gostaríamos de ser, diferente de humano, que é aquilo que somos); “A humanidade não deve passar o resto da eternidade desejando desesperadamente que os programadores tivessem feito algo diferente”; e “Ajudar as pessoas”.

[459] Alguns grupos religiosos colocam uma forte ênfase na fé em oposição à razão, sendo que eles podem considerar essa última — ainda que em sua forma mais idealizada hipoteticamente e mesmo depois de estudar cada escritura, revelação e exegese de forma entusiasmada e de mente aberta — como algo insuficiente para a aquisição de entendimentos espirituais essenciais. Indivíduos que possuem tais pontos de vista podem não considerar a VEC como um guia ideal para a tomada de decisões (embora eles ainda possam preferi-la em detrimento de diversos outros guias imperfeitos que poderiam de fato ser seguidos caso a abordagem VEC fosse evitada).

[460] Uma IA que agiria como uma força latente da natureza na regulação das interações humanas tem sido citada como um “Sysop”, um tipo de “sistema operacional” que atuaria em substituição à civilização humana em diversas questões. Ver Yudkowsky (2001).

[461] “Poderia” porque, supondo-se uma vontade extrapolada coerente da humanidade que não desejasse estender consideração moral a essas entidades, talvez seja duvidoso que essas entidades realmente possuíssem status moral (embora, atualmente, pareça muito plausível que elas tenham). “Potencialmente” porque, mesmo que um poder de veto fosse capaz de impedir a dinâmica da VEC de proteger diretamente esses “forasteiros”, ainda haveria a possibilidade de que, a partir de quaisquer regras básicas remanescentes uma vez que a dinâmica inicial tenha sido executada, os indivíduos cujos desejos fossem respeitados e que quisessem proteger o bem-estar de alguns desses forasteiros seriam capazes de negociar com sucesso e obter esse resultado (abrindo mão de alguns de seus próprios recursos). Se isso seria ou não possível pode depender, entre outras coisas, da possibilidade de que o resultado da dinâmica da VEC seja um conjunto de regras básicas que viabilize resoluções negociadas para questões desse tipo (que poderiam exigir cláusulas para resolver problemas estratégicos de negociação).

[462] Indivíduos que contribuíssem positivamente na realização de uma superinteligência segura e benéfica poderiam merecer alguma recompensa especial por seu trabalho, embora devesse ser algo muito menor do que um mandato praticamente exclusivo para determinar a disposição da herança cósmica da humanidade. No entanto, a noção de que todos deveriam receber uma parte igual na nossa base de extrapolação é um belo ponto de Schelling e não deveria ser deixada de lado facilmente. Há, de qualquer forma, uma maneira indireta pela qual a virtude poderia ser recompensada: a saber, a própria VEC poderia especificar que as pessoas boas que se esforçaram em nome da humanidade deveriam ser devidamente reconhecidas. Isso poderia acontecer mesmo sem que tais pessoas recebessem qualquer consideração especial na base de extrapolação, caso — como é fácil imaginar — a nossa VEC endossasse (no sentido de dar ao menos alguma importância) um princípio de recompensa justa.

[463] Bostrom et al. (2013).

[464] Desde que exista algum significado (suficientemente claro) compartilhado que esteja sendo expressado quando fazemos afirmações morais, uma superinteligência deveria ser capaz de descobrir qual é esse significado. E uma vez que afirmações morais sejam “aptas à verdade” (isto é, tenham um caráter proposicional subjacente que possibilite que elas sejam verdadeiras ou falsas), a superinteligência deveria ser capaz de descobrir quais afirmações no formato “Agente X deve agora Φ” são verdadeiras. Ela deveria, pelo menos, ser superior a nós na execução dessa tarefa.

Uma IA que inicialmente não tenha tal capacidade para cognição moral deveria ser capaz de adquiri-la se ela tiver o superpoder de amplificação da inteligência. Uma maneira pela qual a IA poderia fazer isso seria realizando um processo de engenharia reversa do pensamento moral do cérebro humano e, em seguida, implementando um processo semelhante, porém executado de maneira mais rápida, com informações factuais mais precisas e assim por diante.

[465] Uma vez que temos incertezas em relação à metaética, há uma questão em torno do que a IA teria que fazer se as pré-condições para a RM não fossem obtidas. Uma opção seria estipular que a IA se desligasse automaticamente caso atribuísse uma probabilidade suficientemente alta de que o cognitivismo moral seja falso ou de que inexistam verdades morais não relativas adequadas. Alternativamente, poderíamos fazer com que a IA utilizasse alguma abordagem alternativa, como a VEC.

Poderíamos refinar a proposta da RM de modo a tornar mais claro o que deveria ser feito em vários casos ambíguos ou degenerados. Por exemplo, se a teoria do erro for verdadeira (e, por conseguinte, todas as afirmações morais positivas do tipo “Eu agora tenho que Φ” sejam falsas), então, a estratégia de contingência

(por exemplo, o desligamento) seria invocada. Nós poderíamos também especificar o que deveria ser feito caso existissem múltiplas ações viáveis e cada uma delas fosse moralmente correta. Por exemplo, poderíamos dizer que em tais casos a IA deveria executar as (ou uma das) ações admissíveis que a extrapolação coletiva da humanidade teria escolhido. Poderíamos, talvez, estipular o que aconteceria caso a teoria moral verdadeira não empregasse termos como "moralmente correto" em seu vocabulário básico. Por exemplo, uma teoria consequencialista poderia afirmar que algumas ações são melhores do que outras, mas que não há um limiar específico que seja correspondente à noção de que uma ação é "moralmente correta". Nesse caso, poderíamos dizer que se tal teoria está correta, a RM deveria realizar uma das ações viáveis e moralmente melhores, caso exista uma; ou, caso haja um número infinito de ações viáveis, de tal forma que para qualquer ação viável haja uma ação ainda melhor, então talvez a RM devesse escolher uma que seja ao menos astronomicamente melhor do que a melhor ação que qualquer outro humano teria selecionado em uma situação similar se tal ação fosse viável — caso contrário, ela deveria escolher uma ação que fosse ao menos tão boa quanto a melhor ação que um humano teria realizado.

Quando pensamos nas maneiras pelas quais a proposta da RM poderia ser refinada, dois pontos em particular vêm em mente. Em primeiro lugar, poderíamos começar de forma conservadora, utilizando a opção alternativa para cobrir quase todas as contingências e apenas utilizar a opção "moralmente correta" naquelas que sem dúvidas compreendêssemos perfeitamente. Em segundo lugar, poderíamos acrescentar um modulador geral à proposta da RM para que ela fosse "interpretada de maneira caridosa e revisada conforme teríamos feito caso tivéssemos pensado mais cuidadosamente sobre o assunto antes de escrevê-la etc.".

[466] Entre esses termos, "conhecimento" parecer ser o mais prontamente passível a uma análise formal (nos termos da teoria da informação). Entretanto, para representar o que significa para um ser humano saber algo, a IA poderia precisar de um conjunto sofisticado de representações relativas a propriedades psicológicas complexas. Um ser humano não "conhece" todas as informações que estão armazenadas em algum lugar do seu cérebro.

[467] Um indicador de que os termos na VEC são (marginalmente) menos opacos é que analisar a correção moral nos termos como aqueles empregados na VEC seria um progresso filosófico. Na verdade, uma das principais vertentes da metaética — a teoria do observador ideal — se propõe a fazer exatamente isto. Ver, por exemplo, Smith et al. (1989).

[468] Isso requer que confrontemos o problema da incerteza normativa fundamental. É possível demonstrar que nem sempre é apropriado agir de acordo com a teoria moral que tenha a maior probabilidade de ser verdadeira. Pode-se demonstrar também que nem sempre é apropriado realizar a ação que tenha a maior probabilidade de estar correta. Parece ser necessária alguma forma de troca entre probabilidades e "graus de erro" ou severidade das questões que estão em jogo. Para algumas ideias em relação a esse tema, ver Bostrom (2009a).

[469] Seria possível até argumentar que essa é uma condição de adequação para qualquer explicação da noção de retidão moral que dê conta de explicar como uma pessoa comum é capaz de ter alguma noção de certo e errado.

[470] Não é óbvio que seria moralmente correto que nós construíssemos uma IA que implementasse a RM, mesmo assumindo que a *própria IA* sempre agiria moralmente. Talvez fosse inaceitavelmente prepotente ou arrogante de nossa parte construir tal IA (especialmente se levarmos em conta que muitas pessoas possam não aprovar o projeto). Essa questão pode ser parcialmente refinada por meio de pequenos ajustes na proposta da RM. Vamos supor que estipulássemos que a IA apenas deveria agir (para fazer o que lhe seria moralmente correto fazer) se, a princípio, fosse moralmente correto que seus criadores a tenham construído; caso contrário, ela deveria se desligar. É difícil visualizar como poderíamos estar cometendo algum erro moral grave em criar esse tipo de IA, uma vez que, caso fosse errado tê-la criado, a única consequência seria o desligamento imediato da IA criada, contanto que a IA, até aquele momento, não tivesse cometido nenhum crime da mente. (Poderíamos, contudo, ter agido de maneira equivocada — por exemplo, deixando de aproveitar a ocasião para construir algum outro tipo de IA.)

Uma segunda questão é a supererrogação. Suponha que existissem muitas ações que a IA poderia realizar e que cada uma delas fosse moralmente correta — no sentido de serem *moralmente permissíveis* —, no entanto, algumas seriam moralmente melhores do que as outras. Uma opção é fazer com que a IA tenha como objetivo selecionar a melhor ação em qualquer situação desse tipo (ou uma das melhores ações, caso existam muitas que sejam igualmente boas). Outra opção é fazer com que a IA selecione, entre as ações moralmente permissíveis, uma que satisfaça ao máximo algum outro desiderato (não moral). Por exemplo, a IA poderia escolher, entre as ações moralmente permissíveis, a ação que a nossa VEC preferiria que ela selecionasse. Essa IA não faria nada moralmente inadmissível e provavelmente protegeria melhor os nossos interesses do que uma IA que fizesse o que é moralmente melhor.

[471] Quando a IA fosse avaliar a permissibilidade moral do nosso ato de criação da IA, ela deveria interpretar permissibilidade em seu sentido objetivo. De acordo com um dos sentidos comuns de “moralmente permissível”, uma médica age de modo moralmente permissível quando receita um medicamento que ela acredita ser eficiente em curar seu paciente — mesmo que o paciente, sem o conhecimento da médica, seja alérgico ao medicamento e venha a morrer como resultado disso. Focando na permissibilidade moral objetiva, podemos tirar vantagem da posição epistêmica presumivelmente superior da IA.

[472] E, mais diretamente, dependeria das *crenças* da IA a respeito de qual teoria ética é verdadeira (ou, mais precisamente, de sua distribuição de probabilidade em relação às teorias éticas).

[473] Pode ser difícil imaginar o quanto essas vidas fisicamente possíveis poderiam ser superlativamente maravilhosas. Ver Bostrom (2008c) para uma tentativa poética de fornecer alguma noção de como elas seriam. Ver também Bostrom (2008b) para um argumento a respeito de como algumas dessas possibilidades poderiam ser boas *para nós*, ou seja, para os seres humanos existentes.

[474] Pode parecer enganoso ou manipulador promover uma proposta se julgarmos que alguma outra seria melhor. Mas podemos promovê-la de uma maneira que evite a falsidade. Por exemplo, poderíamos reconhecer abertamente a superioridade de uma proposta ideal e mesmo assim promover uma não ideal como se fosse a melhor proposta factível.

[475] Ou algum outro termo positivamente avaliativo, como “bom”, “ótimo” ou “maravilhoso”.

[476] Isso se parece com um princípio em projeto de software conhecido como *Do What I Mean* [Faça o que eu quero dizer] ou DWIM. Ver Teitelman (1966).

[477] Conteúdo de objetivo, teoria da decisão e epistemologia são três aspectos que deveriam ser elucidados; mas não temos a intenção de levantar a questão a respeito da necessidade ou não de uma decomposição organizada nesses três componentes separados.

[478] Um projeto ético deveria, presumivelmente, alocar no máximo uma parcela modesta dos eventuais benefícios que a superinteligência produzisse como recompensas especiais para aqueles que contribuíssem de formas moralmente admissíveis para o sucesso do projeto. Seria inadequado alocar uma grande proporção ao esquema de pacote de incentivos. Isso seria análogo a uma instituição de caridade que gastasse 90% de sua receita em bônus de desempenho para os funcionários que arrecadaram os fundos e em campanhas de propagandas para aumentar as doações.

[479] Como os mortos poderiam ser recompensados? Podemos pensar em muitas possibilidades. O modo mais simples seria por meio de cerimônias e monumentos, que poderiam ser vistos como recompensas caso as pessoas desejassem notoriedade póstuma. O falecido poderia também ter outras preferências a respeito do futuro que poderiam ser honradas, como algo relacionado a cultura, artes, prédios ou ambientes naturais. Além disso, a maioria das pessoas se importa com seus descendentes, e privilégios especiais poderiam ser concedidos aos filhos e netos dos contribuintes. Indo mais além, a superinteligência poderia ser capaz de criar simulações relativamente fiéis de algumas pessoas do passado — simulações que seriam conscientes e suficientemente parecidas com a pessoa original de modo a contar como uma espécie de sobrevivência (de acordo com os critérios de, pelo menos, algumas pessoas). Isso seria presumivelmente mais fácil para pessoas que tivessem sido colocadas em suspensão criogênica; mas, para uma superinteligência, talvez não fosse impossível recriar algo bastante similar à pessoa original com base em outros registros guardados, tais como correspondências, publicações, materiais audiovisuais e registros digitais, ou a partir das memórias de outros sobreviventes. Uma superinteligência também poderia pensar em algumas possibilidades que não nos ocorrem tão prontamente.

[480] Sobre assaltos de Pascal, ver Bostrom (2009b). Para uma análise de questões relacionadas a utilidades infinitas, ver Bostrom (2011a). Sobre a incerteza normativa fundamental, ver, por exemplo, Bostrom (2009a).

[481] Por exemplo, Price (1991); Joyce (1999); Drescher (2006); Yudkowsky (2010); Dai (2009).

[482] Por exemplo, Bostrom (2009a).

[483] Também é concebível que o uso da normatividade indireta para especificar o conteúdo de objetivo da IA poderia mitigar os problemas que surgiriam com o uso de uma teoria da decisão incorretamente especificada. Considere, por exemplo, a abordagem da VEC. Caso fosse bem implementada, ela poderia ser capaz de compensar pelo menos alguns erros na especificação da teoria da decisão da IA. A implementação poderia permitir que os valores que a vontade extrapolada coerente desejasse que a IA perseguisse dependessem da teoria de decisão da IA. Se nossos egos idealizados soubessem que eles estariam fazendo especificações de valor para uma IA que estivesse usando um tipo particular de teoria da decisão, eles poderiam ajustar suas especificações de valor de modo a fazer com que a IA se comportasse de forma benigna, apesar das distorções em sua teoria da decisão — da mesma maneira como podemos anular distorções em uma lente colocando outra lente na sua frente que gere distorções de modo oposto.

[484] Alguns sistemas epistemológicos podem, seguindo um conceito holístico, não ter uma fundação distinta. Nesse caso, a herança constitucional não é um conjunto distinto de princípios, mas, por assim dizer, um ponto de partida epistêmico que incorpora determinadas propensões para responder à entrada de fluxos de evidência.

[485] Ver, por exemplo, o problema de distorção discutido em Bostrom (2011a).

[486] Por exemplo, uma questão em disputa no campo do raciocínio antrópico é se a chamada suposição da autoindicação deve ser aceita. A suposição da autoindicação afirma, grosso modo, que a partir do fato de que você existe, você deve inferir que hipóteses segundo as quais há uma grande quantidade N de observadores deveriam receber um aumento de probabilidade proporcional a N. Para um argumento contrário a esse princípio, ver o experimento mental do "Filósofo presunçoso" em Bostrom (2002a). Para uma defesa desse princípio, ver Olum (2002); e para uma crítica dessa defesa, consultar Bostrom e Ćirković (2003). Crenças a respeito da suposição de autoindicação podem afetar várias hipóteses empíricas de relevância estratégica potencialmente crucial, por exemplo, considerações como o argumento do juízo final de Carter-Leslie, o argumento da simulação e argumentos do "grande filtro". Ver Bostrom (2002a, 2003a, 2008a); Carter (1983); Ćirković et al. (2010); Hanson (1998d); Leslie (1996); Tegmark e Bostrom (2005). Também poderíamos levantar um ponto semelhante em relação a outras questões preocupantes na teoria da seleção observacional, tais como se a escolha da classe de referência poderia ou não ser relativizada a momentos observacionais e como isso seria possível.

[487] Ver, por exemplo, Howson e Urbach (1993). Há também alguns resultados interessantes que estreitam a gama de situações em que dois agentes bayesianos conseguem discordar racionalmente quando suas opiniões são baseadas em conhecimento geral; ver Aumann (1976) e Hanson (2006).

[488] Cf. o conceito de um "último juiz" em Yudkowsky (2004).

[489] Existem muitas questões importantes pendentes na epistemologia, algumas anteriormente mencionadas no texto. O ponto aqui é que podemos não precisar que todas as soluções estejam rigorosamente corretas para alcançar um resultado que seja praticamente indiscernível do melhor resultado. Um modelo misto (que combina uma gama ampla de diversas funções de probabilidade *a priori*) poderia funcionar.

[490] Princípio introduzido em Bostrom (2009b, p. 190), onde também é observado que ele não é tautológico. Em uma analogia visual, imagine uma caixa grande, mas com volume finito, representando o espaço das capacidades básicas que poderiam ser obtidas por meio de uma tecnologia possível. Imagine que o ato de jogar areia dentro dessa caixa representasse os esforços de pesquisa. Onde você coloca a areia determinará como ela se acumula dentro da caixa. Mas se você continuar a derramar areia, toda a caixa será, eventualmente, preenchida.

[491] Bostrom (2002b).

[492] Essa não é a perspectiva pela qual as políticas de incentivo à ciência têm sido tradicionalmente vistas. Harvey Averch descreveu as políticas científicas e tecnológicas nos Estados Unidos, entre 1945 e 1984, como centradas em debates sobre o nível ótimo de investimento público em empreendimentos de ciência e tecnologia, e também na medida em que o governo deveria tentar "escolher os vencedores" como uma forma de aumentar o máximo possível a prosperidade econômica da nação e seu poderio militar. Segundo esses cálculos, o progresso tecnológico é sempre visto como algo bom, mas Averch também descreve o crescimento de perspectivas mais críticas, que questionam essa premissa de que "o progresso é sempre algo bom" (Averch, 1985). Ver também Graham (1997).

[493] Bostrom (2002b).

[494] Isso, é claro, não é de forma alguma uma tautologia. Seria possível imaginar uma afirmação que advogasse em prol de que essa ordem fosse distinta. Poderíamos argumentar que seria melhor para a humanidade enfrentar um desafio mais simples primeiro, como o desenvolvimento da nanotecnologia, baseando-nos na premissa de que isso nos forçaria a desenvolver instituições melhores, nos tornaria internacionalmente mais coordenados e ajudaria a amadurecer nosso entendimento em relação a estratégias globais. Talvez pudéssemos enfrentar melhor um desafio que nos apresentasse uma ameaça menos metafisicamente confusa do que a imposta pela superinteligência de máquina. A nanotecnologia (ou biologia sintética, ou qualquer que fosse o desafio mais simples que enfrentássemos primeiro) poderia servir como um degrau que nos ajudaria a ascender ao nível de capacidade necessário para enfrentar um desafio de alto nível como a superinteligência de máquina.

Esse argumento teria que ser avaliado caso a caso. Por exemplo, no caso da nanotecnologia, teríamos que considerar várias possíveis consequências, tais como o aumento em desempenho de hardware através do uso de substratos computacionais nanofabricados; os impactos que o capital físico barato para a manufatura causariam no crescimento econômico; a proliferação de tecnologias de vigilância sofisticadas; a possibilidade de que um *singleton* pudesse emergir em virtude de efeitos diretos ou indiretos de uma descoberta tecnológica; e a maior viabilidade das abordagens neuromórficas e da emulação completa do cérebro para a inteligência de

máquina. Não está no escopo de nossa investigação considerar todas essas questões (ou questões paralelas que poderiam surgir em relação a outras tecnologias que seriam capazes de gerar riscos existenciais). Aqui apenas destacamos uma justificativa *prima facie* para favorecer que a superinteligência seja desenvolvida primeiro, ao mesmo tempo que enfatizamos a existência de complicações que poderiam alterar essa avaliação preliminar em alguns casos.

[495] Pinker (2011); Wright (2001).

[496] Poderia ser tentador supor que a hipótese de acelerar tudo o que existe seria inútil nos baseando no fato de que (à primeira vista) isso parece não ter nenhuma consequência observável; no entanto, ver, por exemplo, Shoemaker (1969).

[497] O grau de preparação não é medido pela quantidade de esforços gastos em atividades preparatórias, mas pelo quão favoravelmente configuradas seriam as condições existentes e o quão preparados nossos líderes estariam para tomar as decisões apropriadas.

[498] O grau de confiança entre as nações durante o processo que levará a uma explosão de inteligência também pode ser um fator importante. Isso será levado em consideração mais adiante neste capítulo, na seção “Colaboração”.

[499] Curiosamente, ao que parece, aqueles que hoje em dia estão seriamente interessados no problema do controle compõem uma amostra desproporcional do extremo final da distribuição de inteligência, embora possam existir explicações alternativas para que tenhamos essa impressão. Se esse campo do conhecimento entrar na moda, com certeza será inundado por medíocres e charlatões.

[500] Devo esse termo a Carl Shulman.

[501] Quão similar a um cérebro uma inteligência de máquina precisa ser para que possa ser classificada como uma emulação completa do cérebro, e não uma IA neuromórfica? O fator determinante poderia ser a capacidade do sistema de reproduzir os valores ou a totalidade das tendências cognitivas e avaliativas, quer seja de um indivíduo em particular ou de um ser humano genérico, uma vez que esse aspecto, possivelmente, faria uma diferença em relação ao problema do controle. Capturar essas propriedades pode demandar um grau de fidelidade muito alto nas emulações.

[502] A magnitude desses avanços, dependeria, é claro, do tamanho desses esforços e também da origem dos recursos com os quais estes seriam financiados. Pode não haver nenhum tipo de avanço real em neurociência se todos os recursos direcionados à pesquisa em emulação completa do cérebro fossem deduzidos dos recursos destinados às pesquisas habituais da neurociência — a não ser que, por acaso, acentuar o foco na emulação fosse uma maneira mais efetiva de obter avanços na neurociência do que o portfólio de pesquisas habitualmente realizadas na área.

[503] Ver Drexler (1986, p. 242). Drexler confirmou (por meio de comunicação privada) que essa reconstrução corresponde ao raciocínio que ele buscou apresentar. Obviamente, um número de premissas implícitas deveria ser levado em conta se quiséssemos transformar esse argumento em uma cadeia de raciocínio dedutivamente válida. (Perceba que Drexler não endossa o tipo de argumento relacionado" discutido a seguir.)

[504] Não deveríamos, talvez, dar boas-vindas às *pequenas* catástrofes caso elas aumentassem nossa vigilância a ponto de sermos capazes de prevenir catástrofes de *média escala* que teriam sido necessárias para que tomássemos medidas mais drásticas na prevenção de catástrofes existenciais? (E, é claro, assim como ocorre com sistemas imunológicos biológicos, também devemos nos preocupar com reações muito fortes, semelhantes a alergias e distúrbios autoimunes.)

[505] Cf. Lenman (2000); Burch-Brown (2014).

[506] Cf. Bostrom (2007).

[507] Note que esse argumento está concentrado na ordem, e não na cronologia, dos eventos relevantes. Desenvolver uma superinteligência antecipadamente ajudaria a prevenir outros riscos associados à transição somente se essa intervenção pudesse modificar a sequência de outros avanços cruciais: por exemplo, se a superinteligência surgisse antes de diversos avanços em nanotecnologia ou biologia sintética.

[508] Se a resolução do problema do controle for *extremamente* mais difícil do que a resolução do problema de desempenho em inteligência de máquina e caso exista uma correlação fraca entre a habilidade e o tamanho do projeto, possivelmente seria preferível que um projeto menor fosse bem-sucedido primeiro, já que a variância no fator habilidade é maior entre os projetos menores. Em uma situação como essa, ainda que os projetos menores sejam, em média, menos competentes do que os projetos maiores, poderia ser menos improvável que um projeto menor qualquer viesse a ter o nível surreal de competência necessário para resolver o problema do controle.

[509] Não queremos, com isso, negar que seja possível imaginar a criação de ferramentas capazes de promover deliberações globais que se beneficiariam, ou mesmo necessitariam, de maiores avanços em

hardware — por exemplo, traduções de alta qualidade, melhores ferramentas de busca, acesso universal a smartphones, ambientes de realidade virtual atrativos e propícios à interação social e assim por diante.

[510] Investimentos em tecnologias de emulação poderiam acelerar o progresso em direção à emulação completa do cérebro não apenas de forma direta (através dos avanços técnicos que seriam produzidos), mas também indiretamente, criando um eleitorado que incentivasse esses financiamentos e aumentasse a visibilidade e a credibilidade da visão de uma emulação completa do cérebro (ECC).

[511] Que quantidade esperada de valor poderia ser perdida se o futuro fosse moldado pelos desejos de um ser humano qualquer, e não pelos desejos (com alguma forma de sobreposição adequada) de toda a humanidade? Isso poderia depender sensivelmente do padrão de avaliação aplicado e também se esses desejos em questão fossem idealizados ou puros.

[512] Por exemplo, enquanto as mentes humanas se comunicam lentamente através da linguagem, IAs podem ser desenvolvidas de modo que diferentes instâncias de um mesmo programa sejam capazes de transferir entre si, rápida e facilmente, tanto habilidades como informações. Mentes de máquina desenvolvidas *ab initio* poderiam eliminar os morosos sistemas legados que auxiliaram nossos ancestrais a lidar com aspectos do meio ambiente natural e que não possuem nenhuma importância no ciberespaço. Mentes digitais também poderiam ser desenvolvidas para se aproveitar do processamento serial rápido, indisponível aos cérebros biológicos, o que simplificaria a instalação de novos módulos com funcionalidades altamente otimizadas (por exemplo, processamento simbólico, reconhecimento de padrões, simuladores, mineração de dados e planejamento). Uma inteligência artificial também poderia ter vantagens significativas de caráter não técnico, como ser mais facilmente patenteável ou ficar menos atrelada ao emaranhado de complexidades morais que envolveriam a utilização de emulações humanas.

[513] Se $p1$ e $p2$ são as probabilidades de falha em cada etapa, a probabilidade total de falha é $p1 + (1 - p1)p2$, já que uma falha terminal pode ocorrer apenas uma vez.

[514] Evidentemente, é possível que o projeto líder não tenha uma vantagem tão grande e que não seja capaz de formar um *singleton*. Também é possível que um *singleton* surja antes de uma IA sem a intervenção de uma ECC e, nesse caso, esse argumento em favor de um cenário em que uma ECC surja primeiro perde seu valor.

[515] Existira uma maneira pela qual uma promotora da ECC poderia aumentar a especificidade de seu apoio de modo que o desenvolvimento da ECC acelere ao mesmo tempo que seja minimizado o risco de desenvolvimento acidental de uma IA? Incentivar o desenvolvimento de tecnologias de escaneamento é provavelmente uma aposta mais segura do que incentivar o desenvolvimento de modelagens neurocomputacionais. (É pouco provável que incentivos ao desenvolvimento de hardware venham a fazer muita diferença, em virtude dos grandes interesses comerciais já envolvidos na promoção de avanços nesse campo.)

O fomento de tecnologias de escaneamento poderia aumentar a probabilidade de um resultado multipolar, uma vez que tornaria menos provável que o escaneamento se apresentasse como um gargalo ao desenvolvimento, aumentando, assim, as chances de que as primeiras populações de emulações fossem criadas a partir de diversos modelos humanos, e não apenas compostas de milhões de cópias de um pequeno número de modelos. Avanços em tecnologia de escaneamento também tornariam mais provável que o gargalo se desse em torno do hardware, o que aumentaria a tendência de uma partida mais lenta.

[516] Outros atributos benéficos em termos de segurança presentes na ECC também poderiam estar ausentes em uma IA neuromórfica, tal como um perfil de vantagens e desvantagens cognitivas similar ao de um ser humano biológico (o que nos permitiria empregar nossa experiência com seres humanos na criação de algumas expectativas a respeito das capacidades do sistema durante seus diferentes estágios de desenvolvimento).

[517] Se a motivação de se promover a ECC for a tentativa de fazê-la surgir antes da IA, deve-se ter em mente que a aceleração das pesquisas em ECC alterará a ordem do que surgirá primeiro somente se o tempo padrão dos dois caminhos em direção à inteligência de máquina for próximo e se houver uma pequena vantagem para a IA. Caso contrário, *ou* o investimento em ECC simplesmente resultará no seu surgimento mais cedo do que aconteceria na ausência desse investimento (reduzindo o efeito do excedente de hardware e o tempo de preparação), mas sem afetar a sequência de desenvolvimentos; *ou então* tal investimento em ECC terá pouco efeito (a não ser, talvez, fazer com que, ao estimular o progresso em relação à IA neuromórfica, a IA surja ainda mais cedo).

[518] Comentário feito em Hanson (2009).

[519] Existiria, é lógico, *alguma* magnitude e iminência de risco existencial que justificariam, mesmo a partir da perspectiva do indivíduo-afetado, adiar esse risco — para permitir que as pessoas existentes sobrevivessem um pouco mais antes que as cortinas caíssem ou mesmo para garantir mais tempo aos esforços mitigatórios que poderiam diminuir esse perigo.

[520] Suponha que pudesse ser tomada alguma ação que possibilitasse a ocorrência da explosão de inteligência um ano mais cedo. Consideremos que as pessoas que atualmente habitam o planeta estejam

morrendo a uma taxa de 1% ao ano, e que o risco padrão de extinção da humanidade associado a uma explosão de inteligência seja de 20% (usando aqui um número arbitrário, apenas a título de ilustração). Desse modo, acelerar a explosão de inteligência em um ano pode compensar (do ponto de vista do indivíduo-afetado) um aumento no risco de 20% para 21%, ou seja, um aumento de 5% no nível de risco. No entanto, a grande maioria das pessoas que estivessem vivas um ano antes do início da explosão de inteligência teria, àquela altura, interesse em adiá-la se elas pudessem, dessa forma, reduzindo o risco de uma explosão de inteligência em um ponto percentual (uma vez que a maioria dos indivíduos consideraria que seu próprio risco de morrer no ano seguinte seria muito menor do que 1% — pois a maior taxa de mortalidade ocorre dentro de um quadro demográfico estreito, como os doentes e idosos). Poderíamos ter, então, um modelo no qual, a cada ano, a população votaria pelo adiamento da explosão de inteligência por mais um ano, de modo que essa explosão de inteligência nunca ocorresse, ainda que todos aqueles que estivessem vivos concordassem que seria melhor que uma explosão de inteligência se desse em algum momento. É claro que, na realidade, falhas de coordenação, uma previsibilidade limitada ou preferências outras que não a própria sobrevivência provavelmente poderiam evitar essa pausa sem fim.

Se, em vez da perspectiva do indivíduo-afetado, aplicarmos o fator de desconto econômico padrão, a magnitude da vantagem potencial diminui, uma vez que o valor atribuído à possibilidade de que as pessoas existentes venham a gozar de vidas astronomicamente longas seria acentuadamente descontado. Esse efeito se apresenta especialmente forte se o fator de desconto for aplicado ao tempo subjetivo de cada indivíduo, e não ao tempo sideral. Caso os benefícios futuros sejam descontados a uma taxa de *x*% ao ano, e o nível geral de risco existencial associado a outros fatores for de *y*% ao ano, então o ponto ótimo para a explosão de inteligência se daria quando o adiamento da explosão por mais um ano produzisse menos que *x* + *y* pontos percentuais de redução do risco existencial associado a uma explosão de inteligência.

[521] O conflito poderia, inclusive, representar um risco existencial, principalmente se for precedido pela introdução de novas tecnologias militares de destruição ou um acúmulo sem precedentes de arsenais bélicos.

[522] Um projeto poderia ter seus integrantes espalhados em um grande número de localizações, colaborando através de canais de comunicação criptografados. Mas essa tática tem fatores favoráveis e desfavoráveis à segurança: enquanto a dispersão geográfica poderia oferecer alguma proteção contra ataques militares, a segurança operacional também seria prejudicada, pois haveria uma maior dificuldade em prevenir que membros da própria equipe desertassem, vazassem informações ou fossem sequestrados por rivais se eles estivessem dispersos geograficamente.

[523] O título original desse quadro é “A risk-race to the bottom”. Em inglês, a expressão *race to the bottom* (literalmente, “corrida para baixo”) é usualmente aplicada em linguagem econômica para determinar um quadro competitivo no qual há uma progressiva diminuição e deterioração dos padrões (salários, segurança, excesso de horas trabalhadas) diante de um aumento da competitividade. Aqui escolhemos usar o termo “nivelamento por baixo” para designar essa situação. [NT]

[524] Tenho aqui uma dívida com Carl Shulman e Stuart Armstrong por sua ajuda na elaboração desse modelo. Ver também Shulman (2010a, p. 3): “Chalmers (2010) relata um consenso entre cadetes e funcionários da academia militar norte-americana de West Point, de que o governo americano não restringiria as pesquisas em IA mesmo diante de uma potencial catástrofe, com receio de que potências rivais pudessem ganhar uma vantagem decisiva”.

[525] Note que um acentuado fator de desconto temporal seria capaz de fazer um projeto se comportar como se estivesse inserido em uma dinâmica competitiva, mesmo que fosse conhecida a inexistência de um competidor real. O fator de desconto acentuado diminuiria o peso que o projeto poderia atribuir aos acontecimentos que teriam lugar num futuro distante. Dependendo da situação, isso desencorajaria pesquisas sem uma aplicação clara, o que criaria uma tendência em atrasar a revolução da inteligência de máquina (embora talvez pudesse tornar o seu surgimento algo mais abrupto em virtude do excedente de hardware). Mas o fator de desconto acentuado — ou um baixo índice de preocupação em relação às gerações futuras — também faria com que os riscos existenciais parecessem menos importantes. Isso encorajaria apostas que envolvessem a possibilidade de ganhos imediatos em detrimento de um aumento dos riscos de uma catástrofe existencial, desincentivando, desse modo, investimentos em segurança e incentivando a antecipação do desenvolvimento da inteligência de máquina — reproduzindo os efeitos de uma dinâmica competitiva. Entretanto, ao contrário do que ocorre em uma dinâmica competitiva, um fator de desconto acentuado (ou a desconsideração em relação às gerações futuras) não teria nenhuma tendência em especial de incitar conflitos.

A redução na dinâmica competitiva é um dos principais benefícios da colaboração. O fato de que a colaboração facilitaria o compartilhamento de ideias para a resolução do problema do controle também é um benefício, no entanto, isso seria, em certa medida, compensado pelo fato de que a colaboração também facilitaria o compartilhamento de ideias a respeito da solução do problema da competência. O efeito líquido da facilitação do compartilhamento de ideias pode ser um pequeno aumento da inteligência coletiva da comunidade científica relevante.

[526] Por outro lado, a supervisão pública por um único governo implicaria no risco de produzir um cenário no qual uma única nação monopolizasse todos os ganhos. Essa parece uma situação pior do que um cenário no qual indivíduos altruístas sem responsabilidade formal garantissem que todos pudessem ganhar alguma coisa. Além do mais, a supervisão por parte de um governo nacional não significaria, necessariamente, nem ao menos que todos os cidadãos desse país receberiam uma parte dos benefícios: dependendo do país em questão, existiria um risco maior ou menor de que todos os benefícios fossem capturados por uma elite política ou por alguns funcionários públicos oportunistas.

[527] Uma condição é que a aplicação de um pacote de incentivos (conforme discutido no capítulo 12) poderia, em algumas circunstâncias, encorajar as pessoas a participarem de um projeto como colaboradores ativos em vez de aproveitadores passivos.

[528] Retornos decrescentes, ao que parece, se estabeleceriam numa escala muito menor. A maioria das pessoas preferiria ter uma estrela em vez de uma chance de um em 1 bilhão de possuir uma galáxia com bilhões de estrelas. De fato, a maioria das pessoas optaria por ter acesso a um bilionésimo dos recursos da Terra do que uma chance de um em 1 bilhão de possuir todo o planeta.

[529] Cf. Shulman (2010a).

[530] Teorias éticas agregativas encontram problemas quando a possibilidade de infinitude do cosmos é levada a sério; ver Bostrom (2011b). Também podem haver problemas quando consideramos a ideia de que existem valores ridiculamente altos que podem ser finitos; ver Bostrom (2009b).

[531] Se aumentamos o tamanho de um computador arbitrariamente, em algum ponto encontramos limites relacionados a latências na comunicação entre as diferentes partes do computador — sinais não se propagam em velocidades mais rápidas que a da luz. Se diminuímos o tamanho do computador arbitrariamente, encontramos limites quânticos relativos a essa miniaturização. Se a densidade do computador for aumentada arbitrariamente, o limite seria de um buraco negro. É bem verdade, no entanto, que não temos certeza absoluta de que novas leis físicas não poderão ser descobertas um dia, oferecendo, então, algumas alternativas a essas limitações.

[532] O número de cópias de uma pessoa aumentaria linearmente na ausência de um limite superior para aos recursos. No entanto, não é claro até que ponto um ser humano normal valorizaria o fato de possuir múltiplas cópias de si mesmo. Inclusive as pessoas que preferissem possuir múltiplas instâncias de si poderiam não ter uma função de utilidade que aumentasse linearmente com um número crescente de cópias. O número de cópias, assim como o de anos, pode apresentar retornos decrescentes na função de utilidade de uma pessoa típica.

[533] Um *singleton* é bastante colaborativo internamente em sua mais alta instância decisória. Um *singleton poderia* apresentar falta de colaboração e tendência a conflitos em instâncias inferiores, caso sua instância superior assim preferisse.

[534] Se cada equipe rival de IA estiver convencida de que as outras equipes estão tão erradas a ponto de não terem nenhuma chance de produzir uma explosão de inteligência, então uma das razões para a colaboração — evitar uma dinâmica competitiva — seria eliminada: cada equipe deveria optar de modo independente por seguir em um ritmo mais lento, confiante de que não enfrentaria nenhum competidor relevante.

[535] Um aluno de doutorado.

[536] Pretende-se que a leitura dessa formulação inclua uma prescrição de que o bem-estar de animais não humanos e outros seres conscientes (incluindo mentes digitais), que existem ou venham a existir, seja levado em consideração. Ela não deve ser lida como uma licença para que o desenvolvedor ou desenvolvedora de uma IA substitua as suas próprias instituições morais por aquelas de uma comunidade moral mais ampla. Esse princípio é consistente com a abordagem da “vontade extrapolada coerente”, discutida no capítulo 12, com uma base de extrapolação que engloba todos os seres humanos.

Esclarecendo ainda mais: a formulação não pretende necessariamente excluir a possibilidade da existência de direitos de propriedade de inteligências artificiais, dos algoritmos que as constituem ou de suas estruturas de dados no pós-transição. A formulação deve ser agnóstica a respeito de quais sistemas legais ou políticos seriam mais eficientes na organização das transações em um futuro hipotético de uma sociedade pós-humana. O que essa formulação pretende *é* afirmar que a escolha de tal sistema, na medida em que sua seleção seja casualmente determinada pelo modo de desenvolvimento da superinteligência, deveria ser determinada com base nos critérios aqui descritos; ou seja, o sistema constitucional numa pós-transição deveria ser escolhido em benefício de toda a humanidade e estar a serviço de ideais éticos amplamente compartilhados — e não, por exemplo, meramente em benefício de quem quer que desenvolva primeiro uma superinteligência.

[537] Obviamente, é possível refinar essa cláusula de lucro inesperado. Por exemplo, o limiar talvez pudesse ser expresso em termos *per capita*, ou talvez devêssemos garantir que o vencedor pudesse manter para si uma parte um pouco maior do que uma parcela igualitária do excedente como forma de incentivar uma produção

mais elevada (a adoção de algo como uma versão do princípio maximin de Rawls poderia ser algo atrativo aqui). Outros aperfeiçoamentos poderiam consistir em desvincular a cláusula de valores em dólares e reescrevê-la em termos da "influência no futuro da humanidade" ou "o grau pelo qual o interesse de diferentes partes será ponderado na função de utilidade de um *singleton* futuro" ou algo que o valha.

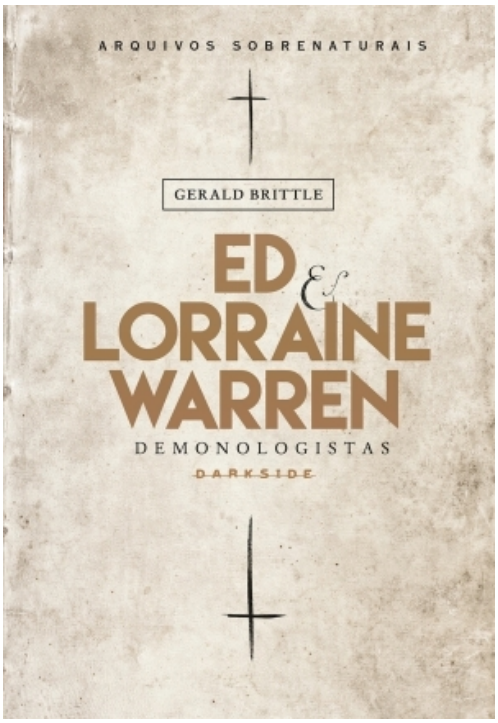
[538] Algumas pesquisas valem a pena não apenas pelas descobertas que proporcionam, mas também por outras razões, tais como entreter, educar, credenciar ou encorajar aqueles que estejam envolvidos nela.

[539] Não estou sugerindo que *ninguém* deveria trabalhar em matemática pura ou filosofia. Também não estou sugerindo que esses esforços sejam um desperdício maior do que todos os outros desperdícios do meio acadêmico ou da sociedade em geral. É provavelmente muito bom que algumas pessoas possam se dedicar a uma vida em prol da mente e seguir a sua curiosidade intelectual para onde quer que ela as leve, independentemente de qualquer utilidade ou eventual impacto que isso possa causar. A sugestão é que, na margem, algumas das melhores mentes poderiam, ao perceber que seus desempenhos cognitivos possivelmente se tornariam obsoletos num futuro próximo, desejar redirecionar sua atenção aos problemas teóricos cuja solução antecipada faria diferença.

[540] Embora deva haver cautela nos casos em que essa incerteza possa se apresentar como protetora — lembre-se, por exemplo, do modelo de nivelamento por baixo apresentado no quadro 13, no qual concluímos que informação estratégica adicional poderia ser prejudicial. De modo mais geral, precisamos nos preocupar com perigos relacionados à informação (ver Bostrom [2011b]). É tentador afirmar que são necessárias mais análises a respeito dos perigos informacionais. Isso é provavelmente verdadeiro, mas ainda deveríamos temer a possibilidade de que essas próprias análises produzam informações perigosas.

[541] Cf. Bostrom (2007).

[542] Sou grato a Carl Shulman por enfatizar esse ponto.


ARQUIVOS SOBRENATURAIS
GERALD BRITTLE
ED & LORRAINE WARREN
DEMONOLOGISTAS
DARKSIDE

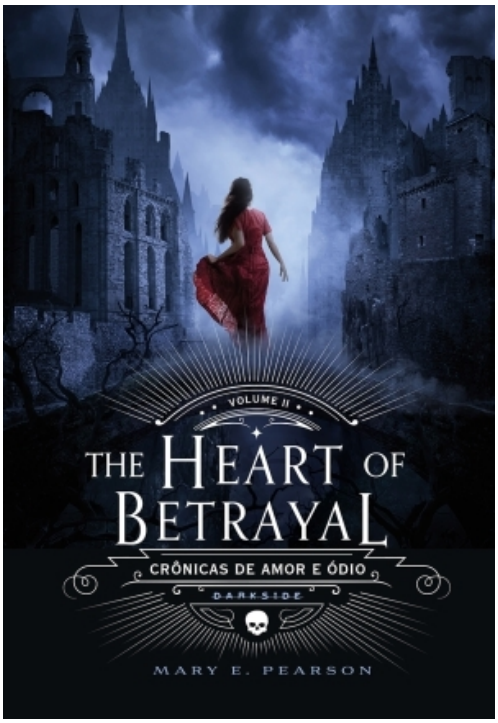
Ed & Lorraine Warren: Demonologistas

Brittle, Gerald
9788594540645
272 páginas

Compre agora e leia

Eles enfrentaram os mistérios mais sinistros dos últimos sessenta anos, sempre em busca da verdade. Agora é a sua vez de entrar em contato com o sobrenatural. Você tem coragem? Então leia Ed & Lorraine Warren: Demonologistas, a biografia definitiva dos mais famosos investigadores paranormais do nosso plano astral. Não é de hoje que os fãs do terror conhecem Ed Warren e sua esposa, Lorraine. O casal foi retratado em filmes de grande sucesso, como Invocação do Mal, Annabelle e Horror em Amityville. Mas basta folhear as páginas de Ed & Lorraine Warren: Demonologistas para constatar que, muitas vezes, a vida pode ser bem mais assustadora que o cinema. No livro, Gerald Brittle desvenda alguns dos principais casos reais vividos pelos Warren. Ed e Lorraine permitiram ao autor acesso exclusivo aos seus arquivos sobrenaturais, que incluem relatos extraordinários de poltergeists, casas mal-assombradas e possessões demoníacas. O resultado é um livro rico em detalhes como nenhum outro.

Compre agora e leia


VOLUME II
THE HEART OF BETRAYAL
CRÔNICAS DE AMOR E ÓDIO
DARKSIDE
MARY E. PEARSON

The Heart of Betrayal

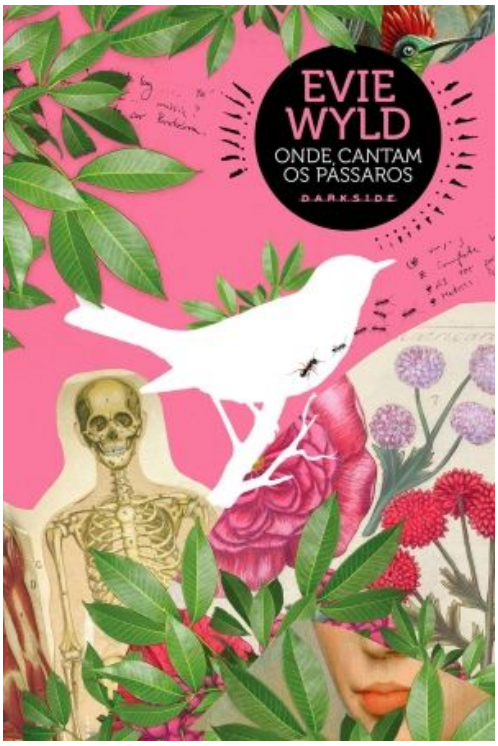
Pearson, Mary E.
9788594540560
400 páginas

Compre agora e leia

A série Crônicas de Amor e Ódio, iniciada com THE KISS OF DECEPTION, virou a queridinha dos leitores mais apaixonados. Encantou os fãs de fantasia do mundo todo - e pegou os brasileiros pelo coração. A história de Lia inspirou muitos leitores a embarcarem em uma jornada extraordinária repleta de ação, romance, mistério e rivalidade, pintados em um universo deslumbrante criado pela premiada escritora Mary E. Pearson, que consegue - como poucos - erguer um mundo poderoso e repleto de personagens cativantes. Em The Heart of Betrayal, Lia e Rafe estão presos no reino barbárico de Venda e têm poucas chances de escapar. Desesperado para salvar a vida da princesa, Kaden revelou ao Vendan Komizar que Lia tem um dom poderoso, fazendo crescer o interesse do Komizar por ela. Enquanto isso, as linhas de amor e ódio vão se definindo. Todos mentiram. Rafe, Kaden e Lia esconderam segredos, mas a bondade ainda habita o coração até dos personagens mais sombrios. E os Vendans, que Lia sempre pensou serem selvagens, desconstroem os preconceitos da princesa, que agora cria uma aliança inesperada com eles. Lutando com sua alta educação, seu dom e sua percepção sobre si mesma, Lia precisa fazer escolhas poderosas que vão afetar profundamente sua família... e seu próprio destino. Não faltam reviravoltas nessa incrível fantasia que integra a linha Darklove - Mary E. Pearson vai deixar os leitores sem fôlego nesse segundo ato de uma

trilogia que ainda promete arrebatar o coração de muitos. Empolgada com a receptividade calorosa dos brasileiros, Mary está se sentindo em casa com toda a repercussão de sua trilogia por aqui: "Nunca imaginei ter tantos fãs no Brasil! Espero visitar vocês em breve".

Compre agora e leia


EVIE
WYLD
ONDE CANTAM
OS PÁSSAROS
DARKSIDE

Onde cantam os pássaros

Wyld, Evie
9788594540836
256 páginas

Compre agora e leia

No premiado romance de Evie Wyld, a fazendeira Jake White leva uma vida simples numa ilha inglesa. Suas únicas companhias são rochedos, a chuva incessante, suas ovelhas e um cachorro, que atende pelo nome de Cão. Tendo escolhido a solidão por vontade própria, Jake precisa lidar com acontecimentos recentes que põem em dúvida o quanto ela realmente está sozinha – e o quanto estará segura. De tempos em tempos, uma de suas ovelhas aparece morta, o que pode ser muito bem obra das raposas que habitam a floresta próxima à sua fazenda. Ou de algo pior. Um menino perdido, um homem estranho, rumores sobre uma fera e fantasmas do seu próprio passado atormentam a vida de uma mulher que sonha com a redenção. Aos poucos, vamos descobrindo mais sobre as suas habilidades em tosquiar e cuidar de ovelhas, aprendidas ainda quando jovem, em sua terra natal, na Austrália. E vamos aprendendo também o que aconteceu lá, que acabou por conduzir White à uma vida de reclusão e isolamento. E sobre as contradições e diferenças entre um passado (sempre narrado no tempo verbal presente) cheio de vida e calor, e o presente (narrado por sua vez no passado) repleto de lama, frio e um ritmo mais desacelerado, paira uma atmosfera absolutamente brutal. Com uma prosa verdadeiramente excepcional, o estilo da autora reúne tanto clareza como substância e apresenta uma personagem inesquecível, enigmática, trágica, assombrada por

um passado inescapável. Uma mulher forte, ainda que tão passível de falhas, erros e equívocos como todos nós. É uma história de solidão e sobrevivência, culpa, perda e o poder do perdão. Uma escrita visceral onde sentimos a presença de tudo, os odores, o vento, o tempo. Nada passa desapercebido.

Compre agora e leia



Donnie Darko

Kelly, Richard
9788594540812
240 páginas

Compre agora e leia

Você ainda não viu esse filme. Mesmo que seja fã de carteirinha, e já tenha assistido a Donnie Darko mais de uma vez – ou dez, ou quinze, quem está contando? Pois a versão que a DarkSide Books está lançando é inédita. Pela primeira vez no Brasil, você poderá entrar de cabeça no universo paralelo de Frank, Donnie e seu criador. Não se trata de uma adaptação do longa-metragem de 2001, e muito menos de um romance que poderia ter inspirado o cineasta Richard Kelly. Donnie Darko, o livro, apresenta na íntegra o roteiro original. A primeira materialização da história, sua chance de conhecer a visão original dessa intrigante obra-prima. A história de Donnie Darko é fácil de resumir – e talvez por isso mesmo ele tenha se transformado no filme preferido de tanta gente mundo afora: um adolescente problemático, com sintomas de esquizofrenia e sonambulismo, escapa da morte quando uma turbina de avião cai no seu quarto. Ele passa a ter visões com Frank, o humano numa estúpida roupa de coelho, ou seria o contrário? Além do roteiro original, Donnie Darko, o livro conta com prefácio exclusivo, assinado por Jake Gyllenhaal, o astro de sucessos como Zodíaco e O Segredo de Brokeback Mountain; uma robusta entrevista sobre todo seu processo de criação, verdadeira aula sobre o amor ao cinema e as armadilhas da indústria do entretenimento; e A Filosofia da Viagem no Tempo – isso mesmo, uma reprodução de trechos do livro escrito por Roberta Sparrow, a Vovó Morte do

filme. É o livro que Donnie lê para tentar desvendar o que está acontecendo no mundo ao seu redor. Agora você tem a mesma oportunidade.

Compre agora e leia


A 2ª FASE DO JOGO.
TRILOGIA
THE GAME
VOLUME 2
ANDERS DE LA MOTTE
DARKSIDE

The game: Ruído

De La Motte, Anders
9788566636727
312 páginas

Compre agora e leia

Como acontece num bom videogame, à medida que o jogo avança, a leitura fica mais perigosa. RUÍDO, volume dois da Trilogia The Game, traz o protagonista HP Peterson enfrentando uma nova fase do Jogo de Realidade Alterada que pôs sua vida em risco, no primeiro livro da série. HP poderia ter tudo: dinheiro, conforto, liberdade. Mas ele está disposto a arriscar tudo para sentir de novo a adrenalina correndo em suas veias. Enquanto isso, a policial Rebeca Normén começa a receber ameaças anônimas por um fórum de internet. O cerco começa a se fechar sobre os dois. Como se proteger de uma ameaça que você não tem certeza que existe. RUÍDO é o segundo livro da Trilogia The Game, de Anders de la Motte, o ex-policial e diretor de segurança de informação que se transformou no grande nome do suspense da Suécia após a morte de Stieg Larsson. O autor desenvolve uma série para a TV americana com duas produtoras: Gaumont (de Narcos) e a dinamarquesa Good Company Films (que adaptou os livros de Stieg Larsson). A Trilogia The Game conta a história de HP, um jovem que tem sua vida transformada num jogo emocionante quando encontra um celular no vagão de trem. Através de mensagens anônimas no aparelho, ele passa a receber instruções para realizar tarefas no mínimo instigantes. A detetive Rebecca Normén é sua irmã, diferente de HP como são opostos a água e o vinho. Fenômeno em diversos países, a Trilogia The Game é surpreendente,

divertida e assustadora na medida certa. Um thriller dos tempos de hoje, onde tudo o que acontece numa tela touchscreen já não pode mais ser considerado virtual.

Compre agora e leia

Table of Contents

Mídias sociais
Folha de rosto
Sumário
Figuras, tabelas e quadros
Prefácio
Capítulo 1
- Desenvolvimentos anteriores e capacidades atuais
  - Regimes de crescimento e a história
  - Grandes expectativas
  - Períodos de esperança e aflição
  - Estado da arte
  - Opiniões sobre o futuro da inteligência de máquina

Capítulo 2
- Caminhos para a superinteligência
  - Inteligência artificial
  - Emulação completa do cérebro
  - Cognição biológica
  - Interfaces cérebro-computador
  - Redes e organizações
  - Resumo

Capítulo 03
- Formas desuperinteligência
  - Superinteligência rápida
  - Superinteligência coletiva
  - Superinteligência de qualidade
  - Alcance direto e indireto
  - Vantagens da inteligência digital

Capítulo 04
- A cinética de uma explosão de inteligência
  - Momento e velocidade da partida
  - Resistência
    - Caminhos de inteligência de não máquinas
    - Caminhos de emulação e IA
  - Poder de otimização e explosividade

Capítulo 05
- Vantagem estratégica decisiva
  - O projeto líder terá uma vantagem estratégica decisiva?
  - Qual será o tamanho de um projeto bem-sucedido?
    - Monitoramento
    - Colaboração internacional
  - De uma vantagem estratégica decisiva para o singleton

Capítulo 06
- Superpoderes cognitivos
  - Funcionalidades e superpoderes
  - Um cenário de tomada do poder por uma IA
  - Poder sobre a natureza e os agentes

Capítulo 07
- A vontade superinteligente
  - A relação entre inteligência e motivação
  - Convergência instrumental
    - Autopreservação
    - Integridade do conteúdo do objetivo
    - Aprimoramento cognitivo
    - Perfeição tecnológica
    - Aquisição de recursos

Capítulo 08
- O resultado mais provável será o nosso fim?
  - Catástrofe existencial como resultado esperado de uma explosão de inteligência?
  - A curva traiçoeira
  - Modos de falha malignos
    - Instanciação perversa
    - Profusão de infraestrutura
    - Crimes da mente

Capítulo 09
- O problema do controle
  - Dois problemas de agência
  - Métodos de controle de capacidade
    - Métodos de confinamento
    - Métodos de incentivo
    - Inibição
    - Detonadores
  - Métodos de seleção de motivação
    - Especificação direta
    - Domesticidade
    - Normatividade indireta
    - Ampliação
  - Sinopse

Capítulo 10
- Oráculos, gênios, soberanos e ferramentas
  - Oráculos
  - Gênios e soberanos
  - IAs ferramentas
  - Comparação

Capítulo 11
- Cenários multipolares

Sobre cavalos e homens
Salários e desemprego
Capital e assistência social
O princípio malthusiano em uma perspectiva histórica
Crescimento populacional e investimento
A vida em uma economia algorítmica
Escravidão voluntária, morte casual
Trabalhar da forma mais eficiente possível seria divertido?
Inconscientes terceirizados?
Evolução não é necessariamente para cima
Formação de um singleton no pós-transição?
Uma segunda transição
Superorganismos e economias de escala
Unificação por tratado
Capítulo 12
Adquirindo valores
O problema da inserção de valores
Seleção evolutiva
Aprendizagem por reforço
Acréscimo associativo de valores
Estepe motivacional
Aprendizado de valores
Modulação de emulação
Projeto institucional
Sinopse
Capítulo 13
Elegendo os critérios de escolha
A necessidade da normatividade indireta
Vontade extrapolada coerente
Algumas explicações
Argumentos para a vec
Observações adicionais
Modelos de moralidade
Faça o que eu quero dizer
Lista de componentes
Conteúdo de objetivo
Teoria da decisão
Epistemologia
Ratificação
Chegando perto o suficiente

Capítulo 14
O panorama estratégico
Estratégia científica e tecnológica
Desenvolvimento tecnológico diferencial
Ordem preferível de chegada
Taxas de mudança e aperfeiçoamento cognitivo
Acoplamento de tecnologias
Antecipando reações
Caminhos e facilitadores
Os efeitos dos avanços em hardware
As pesquisas em emulação completa do cérebro deveriam ser promovidas?
A perspectiva do indivíduo-afetado favorece uma maior velocidade
Colaboração
A dinâmica competitiva e seus perigos
Sobre os benefícios da colaboração
Trabalhando em conjunto
Capítulo 15
A hora decisiva
Filosofia com prazo
O que deve ser feito?
Buscando a iluminação estratégica
Construindo uma boa capacidade
Medidas específicas
Que o melhor da natureza humana se coloque, por favor, de pé
Posfácio
Glossário parcial
Bibliografia
Agradecimentos
Sobre o autor
Créditos