Philosophy and Politics - Critical Explorations

Domingo García-Marzá
Patrici Calvo

Algorithmic Democracy

A Critical Perspective Based on Deliberative Democracy

Springer

Volume 29

Philosophy and Politics - Critical Explorations

Series Editors
David M. Rasmussen
Boston College, Chestnut Hill, MA, USA

Alessandro Ferrara
Dipartimento di Storia, University of Rome 'Tor Vergata', Rome, Italy

Editorial Board
Abdullah An-Na'im
Charles Howard Candler Professor of Law, Emory University, Atlanta, USA

Bruce Ackerman
Sterling Professor of Law, Yale University, New Haven, CT, USA

Robert Audi
O'Brien Professor of Philosophy at the University of Notre Dame, Notre Dame, IN, USA

Seyla Benhabib
Eugene Meyer Professor for Political Science and Philosophy, Yale University, New Haven, CT, USA

Samuel Freeman
Avalon Professor in the Humanities, University of Pennsylvania, Philadelphia, PA, USA

Jürgen Habermas
Professor Emeritus, Goethe-University, Frankfurt am Main, Frankfurt, Bayern, Germany

Axel Honneth
Goethe-University, Frankfurt am Main, Germany and Columbia University, New York, USA

Erin Kelly
Professor of Philosophy, Tufts University, Medford, MA, USA

Charles Larmore
W. Duncan MacMillan Family Professor in the Humanities, Brown University, Providence, RI, USA

Frank Michelman
Professor Emeritus, Harvard University, Cambridge, MA, USA

Tong Shijun
Professor of Philosophy, East China Normal University, Shanghai, China

Charles Taylor
Professor Emeritus, McGill University, Montreal, QC, Canada

Michael Walzer
Professor Emeritus, Institute of Advanced Study, Princeton, NJ, USA

The purpose of Philosophy and Politics - Critical Explorations is to publish high quality volumes that reflect original research pursued at the juncture of philosophy and politics. Over the past 20 years new important areas of inquiry at the crossroads of philosophy and politics have undergone impressive developments or have emerged a new. Among these, new approaches to human rights, transitional justice, religion and politics and especially the challenges of a post-secular society, global justice, public reason, global constitutionalism, multiple democracies, political liberalism and deliberative democracy can be included. Philosophy and Politics - Critical Explorations addresses each and any of these interrelated yet distinct fields as valuable manuscripts and proposal become available, with the aim of both being the forum where single breakthrough studies in one specific subject can be published and at the same time the areas of overlap and the intersecting themes across the various areas can be composed in the coherent image of a highly dynamic disciplinary continent.

Some of the studies published are bold theoretical explorations of one specific theme, and thus primarily addressed to specialists, whereas others are suitable for a broader readership and possibly for wide adoption in graduate courses. The series includes monographs focusing on a specific topic, as well as collections of articles covering a theme or collections of articles by one author. Contributions to this series come from scholars on every continent and from a variety of scholarly orientations.

Domingo García-Marzá and Patrici Calvo

Algorithmic Democracy
A Critical Perspective Based on Deliberative Democracy



Domingo García-Marzá
Universitat Jaume I of Castellón, Castelló de la Plana, Spain

Patrici Calvo
Universitat Jaume I of Castellón, Castellón de la Plana, Spain

ISSN 2352-8370 e-ISSN 2352-8389
Philosophy and Politics - Critical Explorations
ISBN 978-3-031-53014-2 e-ISBN 978-3-031-53015-9
https://doi.org/10.1007/978-3-031-53015-9

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Introduction. How Can We Restrain the Advance of Algorithmic Democracy?

The aim of this book is essentially a modest one. It is intended as an aid for thinking through viable alternatives to the current state of democracy with regard to its ethical foundations and the moral knowledge implicit in or assumed by the way we perceive and understand democracy. It is intended to stimulate reflection and discussion on the basis that, by addressing what we understand as democracy, we can inevitably influence the reality known as democracy. Democracy's evident regression in today's world makes this all too apparent: it has become a hostage to all kinds of autocracies and technopopulisms, which are supported to a greater or lesser extent by the current algorithmic revolution.

The proposal discussed over the course of this book is that the existing concept of democracy, currently faced by the rise of algorithmic democracy, must be extended until it can once again be successfully re-embedded into the contexts of the world we live in. In particular, democracy's institutional embodiment within civil society needs to be reconsolidated. We need to stop confusing democracy and politics; we must revive participation and stop the colonization of civil society by an economy and politics which is structured by macro-data and Big Tech. Our thesis is the following: by thinking in terms of two-way democracy, it is possible to halt the current wave of algorithmic colonization and recover spaces for participation. Civil society and the state can thereby develop all the potential encapsulated by moral autonomy and the civic participation this requires.

The perspective adopted here is a critical one, and as a theory of democracy it breaks with the rigid distinction often drawn between normative theories and empirical theories; it seeks to integrate a range of perspectives and different methodologies to form an ethics of democracy. It is as dogmatic to directly apply moral demands—which is how democracy should work—as it is to deny any form of evaluation (in doing so, logically, one is already evaluating) and impose an empirical horizon as the only explanation for reality.

As is well known, fundamentalism is nothing other than the refusal to justify both one's proposals and methodology, or mode of accessing reality. In short, to think about democracy in accordance with its ethical-political principles simply means proposing arguments for a public debate about what democracy—at least a democracy that strives to respond to expectations of it—might mean. Some of these arguments relate to a moral justification for the democratic principle, and they can be used for the design of political and non-political institutions. Without

this moral core and without the universalist ethics it refers back to, democratic reality ceases to meaningful.

In this book, algorithmic democracy is not seen as a model for democracy, nor does it need to be. In the pages that follow, it is understood as a logical extension of expertocracy: the qualitative leap from knowledge and power ascribing a neutral value to the application of algorithms. With regard to the question of whether we are facing new models of democracy, the answer is no: they are the same models of aggregative, elitist democracy but now supplemented by AI. So it is not a new way of understanding democracy we are facing, but the logical conclusion of the renunciation of equality between people, as well as their moral—and thus also political—autonomy. It is one thing to speak of digital democracy as an instrument or means for participation, whether political or civil, but the progressive and rapid replacement of moral and political autonomy by algorithms is a wholly different matter. The latter scenario has become a reality in many areas of the democratic system: in elections, in decision-making, in shaping public opinion, and so on. We are not talking about complementary new technology serving citizens or individuals. This is a substitution implemented by some citizens, as they might be called, on behalf of the companies that market the algorithms, while also using it to exercise power and control. Over the course of this book, we address the ways in which people have been progressively replaced by data; we term this algorithmic democracy.

If we are incapable of revealing how the marketing and deployment of algorithms follow logics that need to be integrated with practical rationality with the assent of all those potentially affected by them, then the AI technologies will fail to develop in a people-centred way. Instead, this will be the *coup de grâce* for autonomy and, with it, our capacity to decide upon our individual and collective life. Our dignity is threatened by the reductionism implicit in all forms of the datafication encompassing the democratic system, essentially the control and the commercialization of our lives. There is a danger of democracy becoming diluted into an autocracy sustained by the false belief that the future of humanity depends on AI and its algorithms.

Acknowledgements

The contents of this book are part of the objectives of the Scientific Research and Technological Development Project “Cordial Bioethics and Algorithmic Democracy for a Hyperdigitalised Society” [PID2022-139000OB-C22] and “Applied Ethics and Trustworthiness for an Artificial Intelligence” [PID2019-109078RB-C21], funded by the Ministerio de Ciencia e innovación (Ministry of Science and Innovation), as well as the activities of the research group of excellence CIPROM/2021/072 of the Conselleria d’Innovació, Universitats Ciència i Societat Digital (Regional Ministry of Innovation, Universities, Science and Digital Society) of the Valencian Community.

Throughout the development and production of this book, we have received the support of numerous individuals, organisations and institutions. Firstly, we would like to express our sincere thanks to Adela Cortina and Jesús Conill for their contributions and support throughout the process. Secondly, we would like to express our gratitude to the Universitat Jaume I of Castellón for its institutional support, as well as to the Government of Spain, through the Ministry of Science and Innovation, and the Government of the Generalitat Valenciana, through the Regional Ministry of Innovation, Universities, Science and Digital Society, for the generous financial support that has allowed us to carry out this research and to produce and publish this book. Finally, we would like to thank the ÉTNOR Foundation (Foundation for Ethics in Business and Organisations) and its staff for their support.



Contents

Part I Framing Algorithmic Democracy

1 The Democratic Drift

- 1.1 The Democratic Drift
- 1.2 The New Democratic Revisionism
- 1.3 Defining Algorithmic Democracy
- 1.4 Two-Track Democracy
- Bibliography

Part II Ethical and Political Challenges of Algorithmic Democracy

2 The Second Age of Artificial Intelligence

- 2.1 The First Era of Artificial Intelligence
- 2.2 The Second Age of Artificial Intelligence
- 2.3 The Transformative Power of Artificial Intelligence
- 2.4 Ethical and Political Challenges for Democracy
- Bibliography

3 The Virtual Politician: On Algorithm-Based Political Decision-Making

- 3.1 The Paradox of Digitization: The Distrustful Trust in Algorithms
- 3.2 Algorithmic Governance: Virtual Politicians
- 3.3 The Principal Features of Algorithmic Governance
- 3.4 A Critique of Algorithmic Governance: Ethical and Political Issues
- Bibliography

4 Digital Twins: On Algorithm-Based Political Participation

- 4.1 Digital Twins: From Industrial Activity to Political Participation
- 4.2 Augmented Democracy: An Approach Based on Digital Twins
- 4.3 Design Principles and Implementation Requirements for Augmented Democracy
- 4.4 The Ethical and Political Challenges of *Augmented Democracy*

Bibliography

5 Platformization: The Dangers of the Artificial Public Sphere

5.1 The Public Sphere as a Basic Requirement for Democracy

5.2 Platformisation: A New Structural Change in the Public Sphere?

5.3 Cyber-physical Ecosystems, Big Data, and Bots: The Dissolution of the Public Sphere

5.4 Generative Social Bots: The Construction and Development of the Artificial Public Sphere

Bibliography

6 Moral Learning by Algorithms: The Possibility of Developing Morally Intelligent Technology

6.1 The Morality of the Thermostat

6.2 From Artificial Neural Networks to Artificial Moral Intelligence

6.3 Moral Machine Learning Models

6.4 The *Non-place* of the Moral Machine

6.5 Towards the Ethical Governance of Machines

Bibliography

7 The Metaverse: Building a Digital Hyper-economy

7.1 From the Digital Economy to the Artificial Hyper-economy

7.2 Cryptoeconomics: Ethical Aspects of the NFT Revolution

7.3 The Metaverse: The Gateway to the Inner World

7.4 Metanomic Ethics: When the Real Matters

Bibliography

Part III Against Algorithmic Democracy: Scope and Ethical-Discursive Perspectives for an Expansion of Deliberative Democracy

8 Hyperethics: The Automation of Morality

8.1 Deliberation: The Dialogic Transformation of Morality

8.2 Datafication: The Algorithmic Transformation of Morality

8.3 Hyperethics: The Algorithmic colonization of the Lifeworld
8.4 The Ethical and Democratic Challenges Posed by *Hyperethics*
Bibliography
9 A Critique of Algorithmic Reason
9.1 Datapositivism: The New Positivism
9.2 The Ideology of Algorithmic Neutrality
9.3 The Construction of Algorithmic Culture
9.4 Neuropower: From Neuropolitics to Neurodata
9.5 Post-truth: The Power of Technopopulism
Bibliography
10 Dialogic Digital Ethics: From Explicability to Participation
10.1 Towards Dialogic Digital Ethics
10.2 From the Desirable to the Possible
10.3 The Principle of Explicability: From Transparency to the Need for Participation
10.4 Institutional Design: The Role of Applied Ethics in the Democratic System
Bibliography
11 A Critical Civil Society: An Ethical Framework to Control the Power of Algorithms
11.1 Two-Way Democracy: The Value of Participation
11.2 Applied Ethics: Structuring Civil Society
11.3 A Participatory Model of Deliberative Democracy
11.4 Ethical Principles for Institutional Design
Bibliography
12 Institutional Design for the Development of Algorithmic Democracy
12.1 Civil Society Institutions as Agents of Justice
12.2 Ethics and Regulation: Handing Public Engagement with Big Tech

12.3 Corporate Public Sphere: At the Core of Big Tech
12.4 Spheres of Participation and Communication
12.4.1 Ethics Lines and Blockchain
12.4.2 Ethics Lines and Data Mining
12.4.3 Explicability Report
12.5 Spaces for Trust: Recovering Participation
Bibliography
Index

Part I
Framing Algorithmic Democracy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_1

1. The Democratic Drift

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

> *This issue is not just about citizens (this is just one of many ways of being human), but about real people, who are affected by the decisions being taken in the ecological and economic spheres, in the realms of information and technology, in companies and institutions. People are affected in all these places where decisions are made, whereby at the same time, and in keeping with the moral awareness achieved by those societies with liberal democracies, they are valid interlocutors, which means that across these different fields it is necessary to engage not only the perspectives of the expert, and—where appropriate—that of the representative, but also, and above all, that the perspectives of those affected by the decisions made; those who are not mere objects benefiting from them, but autonomous subjects, empowered to and entitled to participate meaningfully in such decisions.*
> *Adela Cortina, Ética aplicada y democracia radical (1993, 19)*

Abstract
The aim of this chapter is to describe and analyse the main causes underlying the emergence and rapid development of what is referred to here as algorithmic democracy. Firstly, we address the reasons why the current mistrust of our liberal democracies, with their commitments to social, economic and environmental rights, is fully deserved, forming part of a phase of increasing disaffection with democracy, which is falling into disrepute. Secondly, we present the principal alternatives to democracy as we know it, founded on populism and technocratic expertocracy. Thirdly, we discuss how the logical outcome of these alternatives is the current digital colonization of all democratic processes and institutions, which we refer to here as algorithmic democracy. Finally, we present a proposal

for a deliberative democracy —a two-way democracy—whose scope of action is not reduced to politics, but rather incorporates the dynamism and power of civil society.

Keywords Algorithms – Digital colonization – Democratic drift – Populism – Democratic revisionism – Two-way democracy

1.1 The Democratic Drift

The speed with which digital colonization has penetrated the heart of democratic institutions and processes has ceased to surprise us. Suddenly, we have seen and felt ourselves living in a digital world in which our moral and political autonomy is increasingly losing its meaning faced with a wave of commercialization and politicization which, of course, includes civil society. In this work we will use the Greek prefix *hyper* with its double sense of surpassing a thing's quality or scale, alluding to excess, two characteristics that define our current digital context. How did we get here? The origin must be sought in the new political scenario: the deterioration and regression that today characterize our democratic systems. Democracy has ceased to be the only desirable system. Moreover, it is losing its meaning and *raison d'être*.

The current degree of social disaffection, our lack of trust in our representatives and their institutions, as well as the principal institutions of civil society, such as companies and universities, are possibly the most important factors to consider to gain an understanding of the rapid introduction of algorithms, along with the emergence and growing importance of what we term algorithmic democracy. "Disaffection" refers to citizens' lack of appreciation and regard for political life, its institutions and representatives. Apathy and privatism are revealed by the lowest ever levels of participation in the major political institutions. Slogans such as "not in my name", or "they do not represent us" lucidly reflect this negative attitude towards politics (Alonso et al. 2011).

Offe (2000, 2011) has identified three basic reasons for the current deterioration of democratic institutions and the lack of credibility and trust generated by this decline, which can potentially lead to the disappearance of all kinds of institution. As an experiment we will apply them to the current situation faced by our liberal democracy and see what results emerge.

The first reason concerns the erosion of the values and rules defining the very meaning of democracy, as well as the institutions that embody them. A good example is provided by the political parties, which have until now been the

principal channel for democratic participation. However, today they have become the opposite: closed systems, dominated by political elites who operate in complete complicity with the economic powers. This results in corruption and a backdoor always open to welcome in the private sector. Meanwhile, parliamentary systems are being hijacked by party discipline and the closed lists of candidates chosen by the leadership. Furthermore, the judicial system has become politicized due to its overseeing bodies and representatives being chosen by politicians. Then there is the media, whose power depends on public budgets and advertising by different governments. The list goes on and on (Mair 2013). In short, we are faced by a real and unequivocal "partitocracy".

When referring to this collusion between the political powers, banks and big business some experts speak of post-democracy (Crouch 2004), while others prefer to speak of extractive elites (Robinson and Acemoglu 2012). Such expressions are undoubtedly exaggerated, and offer little help when thinking about this situation, yet they reflect a widely shared feeling: the disrepute of political parties is undermining the representative system. Citizens neither believe nor trust in their representatives (Levitsky and Ziblatt 2018).

The issue of corruption, a true scourge of our political system and a principal source of distrust, merits special attention. A brief overview of the global situation would clearly show us how this flaw is affecting all levels of administration and, most importantly, how politicians lack the willpower to reverse it beyond approving codes of ethics which, as merely a simple, crude makeover, no one intends to comply with. The perception of corruption increases year on year, yet new control mechanisms are nowhere to be seen (Kubbe and Engelbert 2018). Corruption, in its simplest sense of taking advantage of public office for private gain, raises the question of the value of participation, and this has to be addressed by taking advantage of new digital technologies, in this case the *blockchain*. There is no evidence of a correlation between greater political participation and lower corruption. Worse still, the highest rates of corruption are found in local and regional spheres, where representation is more direct and, theoretically, more effective. Indeed, we know all too well how participation itself can threaten to become corrupt when it is used to gain power, as in the case of so-called political clientelism (Warren 2004). This scenario is bad news for those who, as we will do here, advocate participation as a key factor in meeting the challenges of digital technologies.

The second reason for disaffection is related to the effectiveness of the democratic system, as well as the results expected from the set of institutions making up democracy (Runciman 2018). Now, in addition to a distrust of politicians, increasing frustration has emerged with their policies in the context

of three interlinked crises: the 2008 financial crisis, the 2018 pandemic and the 2021 Russian invasion of Ukraine, which unfortunately still continues. These crises have increased social and economic inequality and have pushed back both the rights that had been achieved and social development. Unlike the crisis of the 1970s, we are not talking about difficulties in meeting the—always high—expectations of the welfare state, but about the failure to meet the minimum standards of justice that guarantee people's dignity and, therefore, democratic citizenship. Growing inequality is the main basis for the current lack of trust (Sen 2009).

According to the *World Inequality Report 2022* (Chancel et al. 2022), there is increasing inequality within these same states. Frustration is giving way to indignation at the current lack of employment—which mainly affects women—and the unacceptably low wages that barely allow people to escape from poverty. The systematic disregard for social and economic rights —the mainstay of civic and political rights— is undermining any attempt at democratic legitimacy. In other words, the social contract that underpinned a decisive part of the credibility of our democracies has deteriorated to such an extent that it is no longer possible to detect any effort being made to attain the equality and social justice expected from democracy. Cuts in social services, job insecurity, together with new forms of exploitation, are leading to distrust in a system that was expected to at least guarantee the minimum conditions for living a life worthy of being described as dignified (Sen 2009). The consequences of neoliberal austerity policies have become much more apparent in the wake of the recent health crisis, which, it should be noted, has forced states to undertake public spending policies. What was deemed impossible during the financial crisis is now economically necessary, which is yet another example of the insincerity underpinning the self-interested exploitation of the supposedly neutral and scientific nature of economic laws (Cortina 2017).

In addition to this social injustice, we must add environmental injustice, which is the source of the current lack of resources. This is one of the challenges which our democracies are failing to meet year after year. As we shall see when discussing the new economies, it is impossible to separate socioeconomic problems from the already irrecoverable state of environmental destruction. We must not fall into the commonplace trap of confusing sustainability with economic efficiency, and the latter with economic rationality. It is absolutely impossible to separate economic, social and ecological sustainability. Today's democracies must take these conditions as a starting point if they want to maintain their legitimacy and thus overcome the current fragility of their institutions (Milanovic 2019). And this same fragility is the main incentive for

the current algorithmic colonization and subsequent search for alternative digital realities.

In the face of these circumstances, two questions arise that demand answers: how much contempt and disaffection is a system capable of withstanding? And how long will it take for the fine thread binding representatives and represented to break? The current distrust of democracy is sweeping away its credibility and leaving it wide open to alternatives that efface its moral core: the recognition of the equality of all people.

Indeed, this is the third reason provided by Offe: the emergence of perceived "better" alternatives to democracy. These include forms of populism fuelled by the failure of liberal and pluralist democracy and, in particular, the outrage prompted by the rapid growth of inequality. Populism is a response to the current disconnection between state and society. It represents a way of thinking, as well as accessing and retaining power, that is not only different from democracy but directly opposed to it, simply because participation and the decision-making concerning everything affecting the formation of the common will ceases to be founded on reason, argumentation, debate and the quest for consensus. Instead, it becomes based on stirring up of emotions, feelings and beliefs; on emotion-provoking and manipulative use of language; on the simplification of reality, and, when necessary, on crassly denying it. In addition, and among other factors, it depends on reducing problems to dilemmas and creating irreducible antagonisms (Mudde and Kaltwasser 2017).

For advocates of populism, democracy must abandon the liberalism of individual rights, the division of powers and representation, and instead move towards a "high-quality democracy" consisting of populist assemblies, with a charismatic, mediatized leadership capable of speaking and acting on behalf of the people as the representative of their true interests against those identified as their enemies and their tactical destruction of patriotic and nationalist sentiments. It also acts against globalization and its elites (Rosanvallon 2020). As we will see below, any form of populism that comes close to expertocracy — both derive from the distrust and lack of credibility of our representative democracies and the disrepute into which they have fallen— can give rise to what we will call technopopulism.

Political parties know very well that it is easier, and more advantageous, to manipulate language and deceive us than to change a clearly unjust reality. In these circumstances, some resort to fear or empty words devoid of commitment. However, it is entirely evident that the first step towards destroying democracy is to deny that there is an intersubjectively valid criterion of justice; a criterion of validity from which to determine any horizon of action other than the current

egotistical mode of “everyone for themselves”. Within this scenario, alternatives emerge, casting doubt for the first time on the ethical roots of democracy, as well as the equality, participation and autonomy underpinning all democratic systems (Cortina 1993).

1.2 The New Democratic Revisionism

The current decline of democracy, evident in both its representative bases and the results of political governance, is blurring the meaning and *raison d’être* of the set of institutions making up the system we refer to as democracy. The recent revisionist thinking is entirely aware of this new scenario, in which, for the first time, democracy is no longer presented as the only or best answer to the question of how to organize our life in common and how to attain a collective formation of will. In this way, a new revisionism has emerged, denying the moral core and ethical foundations sustaining democracy’s significance and institutional development. In this context, democracy risks losing the final vestige of identification between the authors of laws and those they are applied to, and with it the claim that the people affected by the outcome of decisions should be those making them. The spread of populism, digital technologies and the recent vindication of inequality and expertocracy offer a far-from-reassuring prospect. It is therefore necessary to rethink democracy and formulate a critical perspective that enables us to establish a moral foundation for the values that give it meaning, while also highlighting the imminent encircling threats that accompany the new wave of populist revisionism.

The current democratic debate is no longer focused on democratic models (deliberative, elitist, associative, etc.). Instead, it addresses the ethical foundations of democracy, the equality of all those affected by decision-making processes and their inclusion in a range of participatory procedures. Today’s democratic revisionism concurs with the political realism of authors such as Schumpeter (1942) or Sartori (2003), who deny any possibility for a normative or even a moral, dimension of democracy (Joseph 1981). This neorealism has now filtered down to the very foundations of democracy. Doubt is being cast upon it as the best way of organizing life in common, about the basis of equality, and about the moral and political autonomy on which it rests. Alternatives are being presented which, while continuing to invoke democracy, require what are often contradictory predicates. The continuing use of the word democracy, however, even if only to disguise authoritarianism of all kinds, is a sign that it still maintains a power of legitimization, even if this is becoming increasingly vague and thus more susceptible to manipulation and deception.

This is the case with illiberal democracy or democratic authoritarianism, both of which are in fact comparable to many democratic governments today: democracies lacking the basic rights to guarantee equality and participation—in other words, citizens' autonomy. There is no longer any attempt to link intrinsic equality with civic competence (Dahl 1989, 1998). The latter is disputed, the former is denied. The political parties' decline into disrepute is dragging the democratic system down with it. Citizens neither believe nor trust their representatives (Mair 2013). One might even speak of post-democracy if there were a possibility other than authoritarianism and dictatorship, whether that be the state or the markets (Crouch 2004). The liberal foundations of democracy—how to govern—are sacrificed in the name of a democracy—who governs—that revives concepts such as the general will, the common good, participation of the people, and so on, in order to conceal the authoritarianism underlying, for example, the denial of the division of powers or equality in civil and political rights. To deny individual rights, legal protection, the division of powers, representation, political parties, freedom of opinion and the right to assembly is to deny democracy. Illiberal democracies do not exist, just as there is no such thing as an elected dictatorship.

Populism, with its emphasis on emotions, primarily fear; its substitution of argumentative reasons with emotions; its manipulative use of language and revival of charismatic forms of leadership as the true voice of the people, have together set the stage for this revisionist approach and the doubt it casts upon the very roots of democracy. Unlike the early wave of revisionists, their heirs no longer seek to solve these problems through universal suffrage and the election of leaders, but instead aim to disfigure democracy in order to keep it in line with the public lack of interest and the incompetence they perceive (Mounk 2018). Thus, a second type of revisionism sees equal rights and participation being sacrificed in the name of *expertise* (Landemore 2021; Samaržija and Quassim 2023).

A paradigmatic case of this revisionism can be found in Brennan's proposal that democracy has no intrinsic value—no unconditional value—and that we should be open to experimenting with other forms of government whenever necessary. And evidently it has become necessary to do so (Brennan 2016). For this author, democracy is not inherently fair, so we have no moral obligation to cling to inefficient and ineffective institutional structures. Instead we should seek more effective alternatives that achieve better results. He argues that democracy is only a tool, so if we find a better tool, we should feel free to use it. The *epistocracy* Brennan proposes involves the rule of the best educated—the experts—given that, in his view, for the majority of us freedom and political

participation can be detrimental (Brennan 2016). For Brennan, what is important is to achieve good results, and in his view the best resources for decision-making are strategic rationality and economic capacity. Hence the importance of replacing political participation with a market logic and accepting one's lot as a rational ignoramus who neither knows how, nor can, nor even wants to govern (Caplan 2016). The starting point must always be how people perceive themselves: in other words, in terms of their ignorance, irrationality and capacity to be misled. Schumpeter's words in response to Brennan's views on democratic citizenship come to mind: a normal citizen's mental performance drops to a lower level as soon as they enter the field of politics (Schumpeter 1942).

According to Brennan there is no democratic ethos, nor are citizens prepared for or interested in democracy. Moreover, the citizenry is clearly incompetent. Instead there is only a mass of individuals who are driven by selfish motives and emotional impulses. Given this situation, Brennan advocates that we neither can, nor should seek any form democracy beyond that understood as an aggregation of private interests seeking majorities based on utilitarian calculations. He proposes a *government for the people* that is in the interests of the people, under which the satisfaction of our desires and interests is best left to the experts and leaders. Social media undertake to form a consensus for the collective will, defining what the people want and enabling the experts to decide on and implement public policies to achieve these popular goals. Democracy thereby becomes primarily a method for selecting an elite by choosing those deemed most competent. It is therefore argued that the value of the vote ought to depend on each individual's capacity: not all votes are worth the same.

Nevertheless, the recent history of democracy has amply demonstrated how the election of elites has proved to be a failure and always led to *caudillismo* rather than the selection of the best. If democracy is reduced to the addition and subtraction of private interests, akin to a market mechanism in which the consumers are the voters, and the businessmen the politicians, votes are exchanged for services rendered and the experts prove to be as incompetent as the citizens, then representation and voting cease to perform their main function. In this *technopopulism*, which is the outcome of an integration of both forms of political revisionism, political parties, parliament and the professional politicians essentially become ailments debilitating democracy. There is therefore a perceived need to restrict rights and opt for expert-democratic systems in which the most capable and suitable citizens wield greater power. The freedom, equality and participation of all those involved cease to be of importance. They can even become, as we have seen, counterproductive. The naivety of this new revisionism is instantly apparent when we ask ourselves who is going to decide

what efficiency or adequacy means: the people affected or the experts. Saying that more knowledgeable citizens should possess greater power implies defining what is meant by knowledge; what criteria should any hypothetical aptitude test include; and, in addition, who would assess it. Intuitively we know that we have a right to participate in everything that affects us: the concept of dignity is underpinned by this ability. The contrary is nothing other than a disguised form of authoritarianism.

If it is accepted that competence and experience are all that is needed to govern; if citizens merely want their problems solved without caring who or even how this is achieved; and if we have the technological means to undertake this, then the stage is set for the entry of algorithmic democracy.

1.3 Defining Algorithmic Democracy

Having taken the path of replacing those affected by political policies with experts as part of a drive for professionalized, efficient decision-making, based on the latter's supposed neutrality and expertise, it comes as no surprise that amid the digital revolution, with its hyperconnectivity, Big Data and Artificial intelligence, the idea of the algorithmic politician should emerge. This "figure" would make decisions based on billions of pieces of data and never tell lies or intentionally misrepresent information. In other words, it would consist of a set of algorithms as authentic subjects capable of making decisions. Indeed, this is already happening, for example, in some areas of police work and the administration of justice. The algorithmic politician is the latest proposal for creating new democracies based on the objectivity, representativeness and neutrality of mathematical models, which are vaunted as capable of replacing emotionally weak human beings. The latter are considered as the principal cause of bad political decision-making and conflicts of interest, and need to be replaced by quantifiable data on, and the statistical analysis of, the positive aspects of political proposals and citizen petitions, as well as their possible consequences (Matsumoto 2018). There is already talk of *algorithmic democracy*. Neither freedom nor autonomy are of interest here, nor, of course, is responsibility. Needless to say, democracy becomes meaningless as the already slender thread linking the decision-makers—those who construct the algorithms—with those affected by the consequences is severed (Calvo 2019).

Drawing on these premises, we can introduce a concept of algorithmic democracy that spells out the dangers of this digital colonization of democracy:

[Definition]

Algorithmic democracy is understood as a system of social organization and political governance whose scope for action is framed by the gradual incorporation of AI and its technologies in all processes of deliberation, decision-making and institutional design, both in the state and in civil society

Premises:

1. A strictly utilitarian approach to the moral perspective of democracy, whose scope is entirely limited to the calculation of results.
2. An algorithmic culture which allows neither doubt, nor any critical perspective concerning the goodness of digital technologies and how their socio-technical contexts operate.
3. A positivist and scientistic conception of the conversion of feelings, facts, thoughts, and so on into data, and always in an objective and neutral way, from mathematization to the datafication of real situations.
4. The assumption that there is no connection between digital democracy, understood as digitization employed in the service of democratic processes, and algorithmic democracy conceived as the replacement of people with algorithms.
5. The belief that confers on machines equipped with AI the ability to predict human behaviour and make fair, objective decisions based on data.
6. A negative anthropology that reduces the essence of the human being to the satisfaction of particular desires or interests (selfishness or privatism).
7. An instrumentalist and technocratic vision of democracy as being reduced to elections and the representative system and, essentially, to a political aggregation of interests. Democracy is identified with representation and representation with elections, and this clear reductionism equates public responsibility with political responsibility.
8. Subjective rights define a notion of freedom, understood as independence: the ability to do whatever one wants within the permitted scope of action.
9. A mercantilist conception of the public sphere and of civil society in

general, whose power is limited to public opinion and its capacity to influence political decisions.

10. An idea of the economy as a value-neutral hyper-economy that denies any accountability within the democratic system. When confronted by unjust situations, an alternative reality is proposed in which money is the only criterion of validity.

A key difference in situating the set of claims that frame algorithmic democracy is its differentiation from what we commonly refer to as digital democracy. Digital democracy or cyberdemocracy is not an approach, but represents a fundamental and emerging characteristic of twenty-first-century democracy: the use of intelligent methods and technologies to enhance its potential (Dahlberg 2011). However, digital democracy can be either *complementary* or *substitutive* (García-Marzá and Calvo 2022):

1. *Complementary***:** algorithms constitute a type of technique and an emerging technology that is applied as a more efficient, effective and accurate means of helping achieve democratic goals: social order, decision-making and fairer and more desirable public policies.
2. *Substitutive:* algorithms are not only a technique and technology to meet these objectives, they also constitute a horizon of action and are the aim of this technique and technology.

In this second case, human beings, insofar as they view themselves within algorithmic culture, do so as entities who are: a) incapable of making fair decisions due to their characteristic emotional bias; b) being progressively displaced by digitization and its algorithms from taking part in the diverse strata of democratic processes due to their deficient cognitive bandwidth, which includes generating public opinion and electing representatives. Power lies within the algorithms. As citizens concern themselves with living well, algorithms are concerned with recreating the conditions necessary for them to do so: politics, justice, regulatory framework, conditions of possibility and so on (algorithmic paternalism). As a result, within the participatory processes of civil society and politics, human beings are gradually being replaced, which can only result in an *algocracy* where algorithms not only colonize the decision-making processes but also displace human beings from the drafting of legislation, the generation of public opinion, the election of representatives, and so on. In the end, what algo-democracy and algocracy mask is an aristocracy: a democratic

oligopoly. Lying behind the algorithms are powers (individuals, groups of individuals or companies). Algorithmic democracy is the road to algocracy, a meaningless goal devoid of content and an unviable, morally undesirable proposition. There are not and cannot be equal rights between people and machines because machines are not moral agents. They lack the capacity to act freely (they are therefore not responsible for their actions) or establish relationships based on reciprocal recognition.

By invoking the concept of algorithmic democracy, this book does not seek to introduce a new model of democracy, as in our view it is entirely impossible to consider leaving decision-making, public policy, elections and even basic rights in the hands of algorithms. Instead, this concept is intended to highlight the danger implied by progressively replacing participation and autonomy with decision-making and opinion-forming undertaken by data analysis and algorithms. The principal aim of this book is to demonstrate the way that algorithms and their technologies seemingly offer something *complementary* to democracy conceals the *substitution of citizenship.* We are therefore bound to refute the ten points listed above as a form of summary of the premises underpinning today's forms of technopopulism. Our critical perspective towards these premises defines the structure of this book.

The main argument against this new revisionism, with its basis in Artificial intelligence, hyperconnectivity and Big Data, lies in refuting its naïve concept of democracy as a simple technique; an instrument whose meaning begins and ends with how effectively decisions are made, all without invoking any efficiency criterion other than the use of technological power to decide what is fair in the absence of any moral value whatsoever (García-Marzá 2020).

The democratic models underpinning this revisionism are *aggregative* ones, involving the addition and subtraction of particular interests, without any possible discussion of validity or legitimacy of these interests. Compared to them, and as will be seen below, *deliberative models* have a two-fold advantage: they are capable of making the moral core underlying any democratic system explicit, whether it is recognized or not; and they extend the concept of responsibility beyond power. In accordance with *deliberative models*, democracy must be understood in its full breadth and complexity, which in turn demands that the moral dimension inherent in our understanding of democracy should be reconstructed and overseen. As Jürgen Habermas (1992, 1996) and Karl Otto Apel (1988, 1992) have stated, the moral dimension, while normative, has always formed part of the meaning democracy has for us as citizens. The moral sphere is not something outside democracy, it is one of the conditions for making it possible, and it is fundamental to discuss values such as legitimacy or justice,

or, in other words, to maintain a critical perspective.

We must not allow ourselves to be manipulated by wordplay: to substantiate or justify a critical perspective with arguments is not the same as fundamentalism. Indeed, it is the opposite. A fundamentalist is someone who refers to a criterion of validity, such as individualism or privatism, as if it were a natural fact. The same would apply to the neutrality of algorithms or any other social construct. They then refuse to provide justification and simply refer to a pre-constructed reality based on these same theories. Anyone wanting to speak of revisionism from a critical perspective must justify their position, which is otherwise pure and simple dogmatism. Neutrality does not exist in our social world, which is created and maintained through our shared values, ideas and beliefs.

We therefore need to recover democracy's intrinsic value: its moral value. To address this line of enquiry it is important to draw on ideas of discursive ethics and their possible embodiment in a deliberative democracy, a theme we will discuss when we explore the issue of a critical perspective (García-Marzá 2004, 2015). Deliberative democrats who follow this procedural conception of moral validity need to establish a criterion of moral validity—discursive or communicative ethics—capable of demonstrating both the legitimacy of the expectations generated by democracy and the possibility of a set of normative criteria that can be applied as the moral requirement of reciprocal recognition. The latter undertaking would take into consideration our history and the particular situation we live in, which, in the current scenario, means the plural, complex and global contexts shaping our lives. The dual focus on this legitimacy and possibility opens up a critical perspective that can be invoked against the current wave of revisionism, as well as a means of justifying the ethical foundations underlying our trust in democracy and its institutions.

1.4 Two-Track Democracy

Democracy's current drift and the distrust and disaffection that accompanies it potentially risks making it impossible to find any alternative to this new democratic revisionism, whose conceptual framework encompasses algorithmic democracy. Yet we are not only witnessing this burgeoning disaffection, but also the emergence of a strong participatory dynamism. The current lack of interest, commitment and participation in political affairs has been accompanied by the growth of all kinds of initiatives, citizen platforms and movements (Della Porta and Diani 2015); a proliferation of demonstrations and collective protest actions; an increase in the number and strength of solidarity organizations; the creation of

new associations for the defence of common and general interests; and so on. In short, new, innovative forms of participation have arisen outside the structures of the representative system, and this new democratic dynamism is, as will be seen, no stranger to the new digital technologies.

The latest reports provided by *World Protest. A Study of Key Protests Issues in the 21st Century* reveal this to be the case (Ortiz et al. 2021). Last year people took to the streets more than ever before and these demonstrations not only involved activists, trade unionists and social organizations, in particular they also involved health professionals, pensioners, students, women, young people and, in short, a wide range of those bearing the brunt of government policies. It was a powerful reminder that we should stop confusing democracy and representation.

Civil society, which has been lulled to sleep by the misnamed welfare state, is being revitalized by this political activity. Today, participation has become spontaneous and focused on specific causes, with a strong moral commitment. Furthermore, its activity is based in public spaces where personal relationships are still possible. Its actions are anchored within everyday contexts directly related to our interests, our work, our health, our families, and so on (Hirst 2013).

Civil society is not a political society, yet it is a decisive part of democracy. Not all social relations in places like companies, universities, hospitals or churches are political relations, but if they involve power relationships and conflicts occur, these must be solved through a common will: they must form part of a democracy (García-Marzá 2008). It is therefore time to broaden the concept of democracy and to reclaim the value of civil society for the construction of a democracy that contributes meaning and shares responsibility. Is this actually possible in a world reduced to Big Data and algorithms?

From an ethical perspective, democracy cannot be reduced to a model involving the aggregation of any given and unalterable interests that go beyond our will, simply because such an approach disregards our moral autonomy and, therefore, also our capacity to choose a common goal: our capacity to deliberate, to be convinced and to change what is in our interests. Such autonomy expressed in our everyday moral language allows us to discuss misguided interests and to offer an argued critique of the decisions taken by our representatives, even though they may be a majority. A majority does not imply truth, let alone justice, and this becomes entirely apparent when we refer to the injustice of certain aggregations that leave “out” a section of the citizens, or when we consider the inheritance we are bequeathing to future generations. A theory of democracy cannot ignore this moral knowledge; these capacities or resources; this ethical capital.

This is the seed of what is termed a deliberative democracy: a series of theories that understand public deliberation by free and equal citizens as the key to the legitimacy of the political process, and that recognize the existence of a moral core as part of the foundation for all possible legitimation. Likewise, they see democracy, as Carlos Nino would say, as a power for transforming people's interests in a morally acceptable way, and this power is not achieved without the free and equal participation of all those affected (Habermas 1996; Nino 1997; Bohman and Rehg 1999; Elster 2001; Gutmann and Thompson 2004).

Participation is at the very core of this nucleus as a moral requirement derived from autonomy, and it is accompanied by the challenge of understanding this proposal as part of a new participatory democracy. From a moral point of view, this proposal requires the real and effective participation of all those affected, as the only way to discuss justice is with free and equal agreement between all those involved. As soon as the political process ceases to consider preferences and interests as a given, and instead considers them as capable of being discursively transformed, responsibility, or, in other words, the possibility of participating in the transformation of one's own interests, can never be delegated. The participation required is real, direct and effective.

It seems that by getting away from realism we have in fact fallen prey to a crude idealism, not to say a dangerous utopianism. Given the complexity and accelerating dynamics of our societies, seen from the perspective of today's revisionism—just as was the case for realism—any attempt to go beyond our current representative democracy and avoid reducing democratic participation in elections seems naïve, not to say dangerous. Yet this aggregative democracy is precisely the breeding ground of the algorithmic colonization, precisely because of the ease with which our given interests are converted into data and, as will be seen, even produced by algorithms themselves.

Habermas' work on deliberative politics enables us to think through a solution for this situation, one in which the desirable becomes impossible while what is possible is clearly unjust. We consider this important because it allows us to refute the false identification between democracy and politics. For Habermas, democracy is understood as a two-way model: a dual model in which the idea of civil society has to have a prominent place, precisely as the foundation for public opinion and a public discourse devoted to generalizable interests. But it is also as an ethical space, to which each individual makes a contribution.

The methodology underpinning Habermas's proposal is reconstructive and transcendental, among other descriptions, and it consists of making explicit the intuitive knowledge that we, as citizens, possess and put to work when we understand the meaning of democracy and participate in its practices. It is thus a

question of bringing to light a situated mode of reason; knowledge based on the understanding that we, as citizens and participants, have of democracy. If we are capable of accessing this knowledge—what we think democracy should be, and therefore what we expect from it—we will be in possession of a critical standard with which to judge the real practice of democratic reality and thus guide possible alternatives. Theory does not have to describe or prescribe, but it should rather make explicit (reconstruct) the way those involved in and are affected by democratic processes understand their participation.

The outcome of this methodology is an understanding of the political process as a network of discourses and negotiations which include: pragmatic discourses (how to achieve what we want), ethical-political discourses (what we want to be as a collective); negotiations and commitments (agreements between particular interest groups); moral discourses (establishing the criteria of justice for all other discourses) and legal discourses (dedicated to turning the results of the formation of a political will into law). In this sense, for Habermas, the democratic rule of law constitutes the legal institutionalization of this network of discourses and negotiations, the conditions for communication and the procedures that make it possible (Habermas 1992). It is impossible in this brief overview to provide a detailed discussion of the analysis of the form of intuitive knowledge provided by Habermas in *Faktizität und Geltung* (1992), yet three conclusions which affect our understanding of a democracy may be drawn from it.

Firstly, and recalling Kant once again, it may be concluded that there is no politics without ethics. Moral discourses are grafted onto our competence as citizens, and they address the justice of a given practice. They provide a space for a discussion of decisions, rules, commitments, aggregations, and so on, which is always undertaken from the perspective of how interests can be generalized, as well as what we have the right to expect from each other: the *everyone* that includes all the people involved based on their quality as valid interlocutors (Kant VIII).

An example of the presence and meaning of this moral knowledge grafted onto the daily life of democracy is the anti-austerity movement known as the *Indignados*, which revealed how its strength and power of conviction did not depend on the non-fulfilment of any one political programme, but rather on the non-fulfilment of a minimum degree of justice: a minimum that creates the possibility to speak of “dignity”. This is the moral core of democracy, the horizon of action and the argumentative and motivational basis of the participatory dynamism we have already mentioned. Incidentally, this is the case irrespective of the cultural context it operates in.

The second important conclusion to be drawn for a democracy based on

deliberation and agreement addresses a merely instrumental, technical facet of the representative system. In Habermas's view, there is no dichotomy between plebiscitary democracy and representative democracy because the outcome of the network of discourses described must *always* maintain a fallibilist reserve. Participation cannot be delegated, because everyone must deliberate. Any solution reached must therefore always be open to revision and permeable and sensitive to public opinion, which has its own space for action in civil society. In other words, from a moral point of view, respect for "equal participation" requires representative mechanisms (parliament, party competition, universal suffrage, majority rule, and so on) to be open to the influence of public opinion as the voice of generalizable interests (Habermas 1992, 439). Public opinion thus becomes the sounding board that provides the basic problems that have to be managed by the political system. It offers a space where legitimate expectations are configured through communication, and their translation into and realization as policy is monitored and controlled. However, we should not confuse public opinion with published opinion, as we will see in Chap. 5.

The third important conclusion to be drawn is that civil society is needed as the social basis for public opinion. What defines the strength of public opinion is not the fact that it is representative—the quantity of opinions collected through polls, surveys, and so on—but its quality. This quality that corresponds to the way in which it has been produced; that is, its formation process. There is therefore a need for public opinion to have a social space of its own, in which domination-free participation is possible, and where solidarity and understanding are the basic mechanisms for communication; in other words, where communicative action holds sway. In Habermas's view, this is the role of civil society, a space set apart from the sphere of power and money (Bohman 1999, 2000).

With this proposal for two-way democracy, based on the complementary nature of the state and civil society, the participation required for a democratic process is not reduced to political participation, but necessarily extends to all spheres of civil society, and this is clearly demonstrated by the consideration of civil disobedience as a form of democratic participation.

In our opinion, although we can manage to guarantee a critical perspective and to value representation on a technical basis by introducing civil society, in this Habermasian interpretation the sociological basis of public opinion is too far removed from reality: it "excludes" from the concept all those spaces for action and their respective institutions that are not exclusively focused through solidarity and understanding. As a result, the market and the business, along with all other areas involving inequality and asymmetrical power relations, have been

left out of civil society. Yet the source of power lies precisely within these spheres, which are where it is developed. Hence, an ethics for democracy needs to use a broader concept of civil society if it is to deal with the current state of disaffection. To turn to our final point, the value of civil society becomes clear when we see it as part of democracy. It thus provides a solid basis for overcoming the current state of disaffection, which is the outcome of democracy being reduced to political representation. The two-way model of democracy enables it to be rethought as a type of society and not merely as a political regime, in other words, a way of resolving the challenges and conflicts of all coexistence. We will return to this concept of society: one that is not restricted to the formation of public opinion, but instead aims to integrate public responsibility into democracy, as opposed to solely political responsibility.

Deliberative democracies have continued with this idea of a two-way model of democracy. The second generation of authors, such as Bohman (2000) or Dryzek (2006), who began to explore the viability of the initial normative proposals, has been followed by a third generation whose studies on deliberative democracy are markedly descriptive. According to these recent developments, the role of civil society is often summarized in two positions:

1. Micro approaches working with mini-audiences: structured deliberative forums that can engage and “forge links“ with deliberative politics. Participatory financial plans, citizens’ forums and advisory bodies, discussion forums, etc.
2. Macro approaches that tackle the role of civil society as public opinion both “outside” and “confronting” the state, which ranges from social movements to citizen journalism.

In our view, one of the advantages of this so-called “empirical turn” in deliberative democracies has been the discovery of the importance of institutional settings or contexts. The latter discovery shows, for example, that those who claim they are most open to deliberative approaches are those who have withdrawn furthest from partisan policies. However, the same studies also highlight the precariousness of these achievements; they attain only a minimal expression and their importance is beset by a mass of complexities. Furthermore, mention must be made of the elitism implicit in choosing certain subjects for the discussion of certain issues, all within contexts separated from reality (Hendriks 2006).

Yet, in our opinion, undertaking these deliberative proposals does not

achieve the aim of broadening representative democracy, precisely because they continue to accept the sharp division between the system and each individual's respective *lebenswelt* or lifeworld. It is as if companies, or even the medium of money, were not social pacts requiring the context provided by the lifeworld to obtain validity and, therefore, be deemed trustworthy and, in short, have legal authority. By excluding these spheres from their analysis, these authors give visibility to the power exercised by civil society, yet always with reference to the state, as "a second circuit of politics". Such authors do not see the value of civil society in its own right (García-Marzá 2008).

It should be recalled that value means both importance and strength. To counteract disaffection, it is necessary to break out from the dichotomy between communicative action and strategic action and learn to think about which integration of the two logics deserves moral validity. In our opinion, the ethical perspective allows us to escape from this conceptual prison and to make visible the full range of the power of civil society. This is a search for spaces to realize the range of modes characterizing the ultra-social beings are as people, which includes our moral resources. As we will see in the third section of this book, this is the basic contribution made by the *meso-deliberative* approaches to democracy (García-Marzá 2016).

The aim of this book is to use this dual conception of democracy—and we highlight the value of civil society and focus on its institutional structure—as a possible and desirable way to reinstate the digital technology's instrumental character, thereby making it a means for all social activities rather than an end. The strength and power of civil society derives from its wilfulness and spontaneity; from the impossibility of having its borders framed and enclosed within the logic of political or economic power. It is this power that interests us as a means for recovering our capacity for agency in this digital era (García-Marzá 2010).

If we look at the institutions underpinning the different spheres of civil society, we see that they offer a possible foundation for participation and developing autonomy. Institutions are understood here as stable and legitimate social pacts responsible for providing basic social welfare such as peace and security, education, health and efficiency. This is the function performed by schools, universities, hospitals, companies, and so on. Indeed, it is impossible to deny the power that these institutions have and as a result we cannot reject the idea that they play a key role in our democracy and are a part of it. An ethics of democracy can make a small but important contribution to forming a definition of these new spaces for participation in two senses:

– On one hand, ethics can contribute to the design and redesign of institutions: it contributes the basic principles to ensure that all those involved or affected participate within them. Of course, in most cases, representation is required, adjusted to the logic of each social practice, but without the political logic at its core being repeated.
– On the other hand, it is important that the current initiatives emerging from civil society are not dwarfed by the immediacy of the problems confronting them, as these initiatives need to attain diverse degrees of stability. It is neither necessary nor desirable that all the participatory dynamism discussed here should end up being absorbed into political parties. On the contrary, we believe that this mistake has led to the current political colonization: the modern partitocracy.

Power is distributed across a range of social and political spheres. We must learn to manage this power by integrating moral knowledge into the way each of these spheres operates. The fact we can talk about the democratic structure of a hospital does not mean that patients should vote on how surgical operations are to be performed; instead it means that there are spaces for participation within the hospital open to all interest groups. In this way, patients can influence the running of the hospital beyond patient information and care services. As we will see in the third section of this book, the idea behind an ethics of institutions is to establish principles for institutional designs that would enable every stakeholder to participate in each institution (García-Marzá 2020). Public spaces, *within* institutions, for the participation of the those involved and affected, allow institutional decision-making at the heart of institutions and can gather the signals sent out by public opinion. In short, spaces are capable of combining micro- and macro-deliberative approaches. They offer a two-fold possibility for participation aimed at halting the current wave of digital colonialism.

As a result we can broaden the horizon of democracy while respecting the specific nature of each of the spheres it encompasses, but to undertake this task a critical hermeneutic is required: reinstating participation in the lifeworld once more. It is no longer a question of complex, long-term programmes and proposals, but of direct power, within a cultural context, of values and identities. The same people who choose not to campaign for environmental sustainability can do so to prevent an industrial landfill site being created in their municipality. By integrating highly specific practical aims with moral objectives, and issues of general interest with those of collective or even individual interest, our lost vitality can once more be regained.

By adhering to this two-way approach, it becomes feasible to lay claim to the

aspiration of participatory democracy once again. This is firstly because the approach set out here is based on an understanding that political participation is necessary, although it is only one form of participation, which can never replace civil participation. Secondly, and unlike the current deliberative models, this approach understands that participation is not just about deliberation, it is also about deciding, protesting, negotiating, cooperating and so on. In everyday contexts, there is no divide between the language of solidarity and the language of effectiveness. We cannot speak of commitment without talking about "feeling good" and about achieving goals through meaningful, rewarding activities.

Drawing on this diversity of resources and capacities that characterize people, civil society goes far beyond the critical role acknowledged by deliberative approaches. In contrast to Habermas, in the proposal set out here civil society not only exerts an influence, it also wields power (García-Marzá 2016). The principal role of civil society is not to become a right, but rather to involve individuals and their institutions in constructing a democratic will.

Therefore, to conclude this introductory chapter, civil society must learn to use its resources independently to resolve the challenges of joint action. Politics and its institutions are necessary, but they are not sufficient. The complementary relationship between self-regulation and external regulation—between political participation and civic participation—is possible if, and only if, we understand civil society in terms of three core characteristics.

Civil society must be plural, because the spheres of civil society encompass private concerns (prestige, money, etc.), as well as collective concerns (professional, corporate, regulation of common activities, etc.), and general or universal concerns (recognition, dignity, etc.). In addition, in line with the understanding of civil society proposed here, the moral sphere is identified with, but not reduced to, generalizable interests. Communicative action must be integrated with strategic action. The degree of integration attained defines the morality attained and, with it, the capacity to build trust. This is the difference between a good civil society and a bad one.

It must be radical because it includes, firstly, all the institutions that make up this sphere, be they economic, social or cultural; and, secondly, because it includes all those involved in and affected by the activity of these institutions. The criterion of legitimization and, therefore, moral validity—the scope for action—is none other than the possible agreement of all those involved under symmetrical participation conditions. It is solely from this horizon that it is possible to speak of representation.

Nevertheless, this must be comprehensive because it requires participation and agreement regardless of the extent of the problems or the institutions set up

to solve them. The approach therefore sidesteps methodological statism with its reliance on civil society and also deploys its own mechanisms for coordinating action; these are derived from a capacity to establish communicative actions, as well as the moral commitments these imply. The boundaries of civil society—its limits—are determined solely and exclusively by the scale of the conflicts one wishes to solve in a dialogic and reasoned way. This is the meaning of universalism: the recognition of the necessary intercultural dialogue.

Bibliography

Alonso, S., J. Keane, and W. Merkel, eds. 2011. *The Future of Representative Democracy*. Cambridge: Cambridge University Press.

Apel, K.O. 1988. *Diskurs and Veranwortung*. Frankfurt: Suhrkamp.

———. 1992. Diskursethik vor der Problematik von Recht und politik. In *Zur Anwedung der Diskursethik in Politik, recht und Wissenschaft*, ed. K.O. Apel and M. Kettner. Frankfurt: Suhrkamp.

Bohman, J. 2000. *Public Deliberation: Pluralism, Complexity, and Democracy*. Cambridge: The MIT press.

Bohman, J., and W. Rehg, eds. 1999. *Deliberative Democracy*. Cambridge: The MIT Press.

Brennan, J. 2016. *Against Democracy*. Princeton: Princeton University Press.

Calvo, P. 2019. Democracia algorítmica: consideraciones éticas sobre la dataficación de la esfera pública. *Revista del Clad, Reforma y democracia* 74: 5–30.

Caplan, B. 2016. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*. Princeton: Princeton University Press.

Chancel, L., Th. Piketty, S. Saez, G. Zucman. (Coords.) 2022. *World Inequality Report 2022*. World Inequality Lab. https://wir2022.wid.world/www-site/uploads/2022/01/Summary_WorldInequalityReport2022_English.pdf.

Cortina, A. 1993. *Ética aplicada y democracia radical*. Madrid: Alianza.

———. 2017. *Aporofobia, el rechazo al pobre: un desafío para la democracia*. Barcelona: Paidós.

Crouch, C. 2004. *Post-democracy*. Cambridge: Polity.

Dahl, R. 1989. *Democracy and Its Critics*. New Haven: Yale University Press.

———. 1998. *On Democracy*. New Haven: Yale University Press.

Dahlberg, L. 2011. Re-constructing Digital Democracy: An Outline of Four ‘Positions’. *New Media & Society 13* (6): 855–872.

Della Porta, D., and M. Diani, eds. 2015. *The Oxford Handbook of Social Movements*. New York: Oxford University Press.

Dryzek, J.S. 2006. *Deliberative Global Politics: Discourse and Democracy in a Divided World*. Cambridge: Polity.

Elster, J. 2001. *La democracia deliberativa*. Barcelona: Gedisa.

García-Marzá, D. 2004. *Ética empresarial: del diálogo a la confianza*. Madrid: Trotta.

———. 2008. Sociedad civil: una concepción radical. *Recerca: revista de pensament i anàlisi*: 27–46.

———. 2010. El poder de la sociedad civil hacia una ética de las instituciones. *Debats* 109: 30–38.

———. 2012. Kant's Principle of Publicity: The Intrinsic Relationship Between the Two Formulations. *Kant-Studien 103* (1): 96–113.

———. 2015. El valor democrÃtico de la sociedad civil: una respuesta a la desafección. *Themata*. Revista de Filosofiía 52: 93–109. https://doi.org/10.12795/themata.2015.i52.05.

———. 2016. Enfoques mesodeliberativos: sobre la articulación institucional en las democracias deliberativas. *Isegoria* 54: 147–170.

———. 2019. *Habermas: A Two-Track Model of Deliberative Politics*. *Routledge Handbook of Contemporary European Social Movements*: *Protest in Turbulent Times*.

———. 2020. Contra el nuevo revisionismo democrático. El alcance del valor moral de la democracia en K.O.Apel. *Ethic@,* Florianópolis, 19, n. 3, 568–1.

García-Marzá, D., and P. Calvo. 2022. Democracia algorítmica: ¿un nuevo cambio estructural de la opinión pública? *Isegoría* 67: 1–15.

Gutmann, A., and D. Thompson. 2004. *Why Deliberative Democracy?* Princeton: Princeton University Press.

Habermas, J. 1992. *Faktizität und Geltung*. Frankfurt: Suhrkamp.

———. 1996. *Die Einbeziehung des Andere*. Frankfurt: Suhrkamp.

———. 2022. *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Frankfurt: Suhrkamp Verlag.

Hendriks, C.M. 2006. Integrated Deliberation: Reconciling Civil Society's Dual role in Deliberative Democracy. *Political Studies* 54 (3): 486–508.

Hirst, P. 2013. *Associative Democracy: New Forms of Economic and Social Governance*. John Wiley & Sons.

Joseph, L.B. 1981. Democratic Revisionism Revisited. *American Journal of Political Science* 2: 160–187.

Kubbe, I., and A. Engelbert. 2018. Corruption and the Impact of Democracy. *Crime, Law and Social Change* 70 (2): 175–178.

Landemore, Hélène. 2021. An Epistemic Argument for Democracy. In *The Routledge Handbook of Political Epistemology*, ed. Jeroen de Ridder Hannon. London: Routledge.

Levitsky, S., and D. Ziblatt. 2018. *How Democracies Die*. New York: Broadway Books.

Mair, P. 2013. *Ruling the Void: The Hollowing of Western Democracy*. London: Trade Verso.

Matsumoto, T. 2018. *The Day AI Becomes God. The Singularity Will Save Humanity*. Cambridge: Media Tectonics.

Milanovic, B. 2019. *Capitalism, Alone*. Cambridge: Harvard University Press.

Mounk, Y. 2018. *The People vs. Democracy*. Cambridge: Harvard University Press.

Mudde, Rovira. 2017. *Populism. A Very Short Introduction*. Cambridge: Cambridge University Press.

Mudde, C., and C.R. Kaltwasser. 2017. *Populism: A Very Short Introduction*. Oxford University Press.

Nino, C. 1997. *La constitución de la democracia deliberativa*. Barcelona: Gedisa.

Offe, C. 2000. Civil Society and Social Order: Demarcating and Combining Market, State and Community. *European Journal of Sociology/Archives Européennes de Sociologie* 41 (1): 71–94.

———. 2011. Crisis and Innovation of Liberal Democracy: Can Deliberation be Institutionalized? *Sociologický časopis/Czech Sociological Review* 47 (03): 447.

Ortiz, I., S. Burke, M. Berrada, and H.S. Cortés. 2021. *World Protests: A Study of Key Protest Issues in the 21st Century*. Cham: Springer Nature.

Robinson, J.A., and D. Acemoglu. 2012. *Why Nations Fail: The Origins of Power, Prosperity and Poverty*. London: Profile.

Rosanvallon P. 2020. *El siglo del populismo.* Editions du Seuil.

Runciman, D. 2018. *How Democracy Ends*. Profile Books.

Samaržija, H., and C. Quassim. 2023. *The Epistemology of Democracy*. New York: Routledge.

Sartori, G. 2003. *What is Democracy?* Madrid: Taurus.

Schumpeter, J.A. 1942. *Capitalism, Socialism, and Democracy*. New York.

Sen, A. 2009. *The Idea of Justice*. New York: Penguin books.

Warren, M. 2004. What Does Corruption Mean in a Democracy? *American Journal of Political Science* 48 (2): 328–343.

Part II
Ethical and Political Challenges of Algorithmic Democracy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_2

2. The Second Age of Artificial Intelligence

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

Perhaps how wonderful! Think, that for all time, all conflicts are finally evitable. Only the Machines, from now on, are inevitable!
Isaac Asimov, I Robot (1950, 218)

Abstract
The aim of this chapter is to discuss the impact, characteristics and challenges underlying the advent of the Second Age of Artificial Intelligence. This new *hyperdigitized* world offers major possibilities for social development across diverse spheres of activity, such as the economy, health and politics, yet its acritical application is also giving rise to an increasingly complex scenario, making human beings increasingly vulnerable as a result. In this new era for humanity, Artificial Intelligence has seemingly colonized all spheres of human life, making itself the absolute protagonist of progress in all its dimensions, and this has led to the demands placed upon meaning undergoing an exponential growth, which in turn has served as a justification for both the knowledge generated by AI and the actions this can lead to. In the wake of this development there has been an upsurge in the confidence required for the further consolidation and development of AI.

Keywords Algorithms – Algorithmic rationality – Artificial intelligence – Deep learning – Democratic sphere – Digital transformation – Ethics

2.1 The First Era of Artificial Intelligence

The emergence and development of Artificial Intelligence is intrinsically linked to algorithms. Although the design and use of these mathematical models were

common among ancient Babylonian, Egyptian, Greek and Indian civilizations (Banerjee 2021; Blum 2011; Dietel 1987; Infante and Puig 2013; Yuste 2008), the term algorithm was coined in the ninth century by Abu Abdallah Muḥammad ibn Mūsā al-Jwārizmī (Abu Yāffar) in his treatise on algebra, *Compendium of calculus by reintegration and comparison* (*al-Kitāb al-mukhtaṣar fī ḥisāb al-ŷabr wa-l-muqābala*). Derived from the Late Latin term *algobarismus*, and in turn from the Classical Arabic terms *ḥisābu lgubār,* meaning calculation by means of Arabic numerals, Yāffar described algorithms as ordered and finite sequences of algebraic steps or operations that make it possible to find a plausible course of action to resolve a specific dilemma or problem.

Nevertheless, traditional algorithms differ greatly from those currently used in the field of Artificial intelligence. Modern algorithms are linked to the field of computer engineering, and are described as "(…) a finite set of rules that gives a sequence of operations for solving a specific type of problem" (Angius et al. 2021, 4) and they have at least five basic characteristics: finitude, precision, input, output and effectiveness (Angius et al. 2021; Rapaport 2005).

The emergence of the first modern algorithm is closely linked to the work of Ada Lovelace (Hollings et al. 2018). She was fascinated by the machines that were the foundation of the *Industrial Revolution*, such as the loom developed by Joseph Marie Jacquard, which used punched cards to guide its operation, and especially the difference engine developed by Charles Babbage, which was designed to calculate sequences of numbers. Lovelace sought to endow them with greater conceptual elegance and practical application through the use of algorithms (Lovelace 1843). Through her work, Lovelace made significant contributions to the emergence and development of modern algorithms. Especially significant are several of her conceptual contributions (Hammerman and Russell 2016), and worthy of particular note is her major contribution to the design of computer algorithms and the operation of modern computer programs. On one hand, she proposed using Bernoulli numbers—an infinite succession of rational numbers, which are today understood to have a deep-rooted connection with number theory—to describe, using a diagram, the operations Babbage's engine would have to carry out to calculate them. On the other hand, her description of Babbage's engine included ideas which would go on to be developed and widely used in computer programming concepts such as the for loop, which groups together repeatedly executed instructions, and the subroutine, a segment of a program that can be requested at any time in the process. Finally, it was Lovelace who first published a computer program; although Babbage had developed the first computer program, he never circulated it among the scientific community.

Another significant contribution to the emergence of modern algorithms was made by George Boole. In 1854 he proposed a type of algebra based on binary code–Boolean algebra—which was intended to solve problems raised by the application of Babbage's machine. The same idea was used a century later by Alan Turing to build the first modern computers, such as the Turing Machine in 1932, the Oracle Machine in 1937 and 1938, the Enigma Machine in 1940 and the Automatic Computing Engine in 1947. It is still used today to represent data in contemporary intelligent machines (Hammerman and Russell 2016).

Finally, Turing's important contributions to the development of modern computers, software and computational algorithms must be highlighted. One of his best-known works on the subject is "Computing machinery and intelligence" (Turing 1950). In it, Turing showed that it was possible to create computers capable of emulating human intelligence. He also formulated a way of determining whether a machine is sensitive or sentient through a standard test: the Turing test. Moreover, Turing proposed an idea that is still used today in machine learning: it is not necessary to create complex machines that emulate the intellectual capacities of an adult, but simple machines that emulate the learning capacities of a child. This and other studies by Turing, based on the work of Lovelace and Boole, provided a point of departure for the term Artificial Intelligence and its development as a discipline (McCarthy et al. 1955).

The emergence of AI as a scientific discipline is linked to two key moments. Firstly, in 1955 a panel on *Learning Machines* was convened at the Western Joint Computer Conference, which was held in Los Angeles. One of the four papers—inspired by the neuropsychologist Donald Hebb's model for the reinforcement of synapses between neurons—addressed different ways of adjusting the connections within *artificial neural networks* as a model for machines to recognize and learn patterns. Then, at an academic meeting organized by John McCarthy at Dartmouth College in New Hampshire in the summer of 1956, and as stated in the initial call for papers (McCarthy et al. 1955), all the participants committed themselves to making significant contributions to the development of AI, such as significantly improving language comprehension, the abstraction of concepts through learning, and solving of complex problems.

Following its emergence, AI gained followers and supporters and generated high expectations for its development and application. Many of these were based on the idea that, in the short to medium term, it would be possible to achieve *strong AI* that would be capable of adapting to different contexts and making autonomous decisions. As John Searle stated in the 1980s, *strong AI* is AI that is able to adapt to contexts, thinking and making decisions autonomously as a

human being would, while *weak AI* is AI that helps humans carry out mental activities to find optimal solutions to specific tasks. The hypothetical advent of strong AI is what John von Neumann termed technological singularity in 1957 (Ulam 1958, 5).

The emergence of both *weak AI* and *strong AI* –as well as *super AI* or *supra AI*, which are discussed below) depended on different lines of development based on a number of mathematical or biological models, which are rooted in or are, to a greater or lesser extent, compatible with the *physical symbol systems* (PSS) formulated by Allen Newell and Herbert Simon in 1975. The idea underpinning the latter is that it is possible to establish "(…) a set of entities called symbols which, through relationships, can be combined to form larger structures—such as atoms forming molecules—which can be transformed by applying a set of processes" (López de Mántaras and Meseguer 2017, 9).

On the one hand, the *symbolic model*, which was inspired by mathematics and whose application was not necessarily corporeal or linked to a real environment, operates "(…) with abstract representations of the real world that are modelled by means of representational languages based mainly on mathematical logic and its different ramifications" (López de Mántaras and Meseguer 2017, 11).

On the other hand, the *connectionist model* was inspired by biological studies, although its application was not necessarily corporeal or linked to a real environment, took the synaptic activity of brain neurons as its reference point. It was based on an understanding of intelligence as the consequence of the distributed activity of a large number of interconnected units that process data in parallel.

Finally, the *evolutionary model* was once more inspired by advances in biology, but again it was not necessarily corporeal or linked to a real environment. It imitated the process of natural selection—especially the *dynamics of replication* and *evolutionarily stable strategies*—which automatically refines the possible courses of action for the resolution of specific problems.

However, in the mid-1970s, many theorists and practitioners of Artificial intelligence began to realize that both the predictions made by the founders of the discipline and the expectations they had generated over the following three decades were overly optimistic, unachievable, lacking in conceptual rigour or highly unrealistic, and this led to public loss of interest. Subsequently, many of the professionals and researchers involved in both the theoretical development and practical application of AI abandoned their lines of research (López de Mántaras and Meseguer 2017, 21). Among other things, the problems and

limitations, both theoretical and technological that emerged from these three lines of research made it clear that *strong AI* was practically impossible to achieve.

One of its most pressing problems derives from the incompleteness theorem developed by Kurt Gödel, who in 1930 proved that even mathematical systems cannot be complete and self-sufficient. In other words, there are propositions that, although true, are not demonstrable, which implies that any attempt to deductively infer the complete reality of the world from certain axioms results in inconsistency (Calvo 2018, 2022, 2024; Larson 2021, 18–21; Goldstein 2005). As Roger Penrose argues in The *Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (1989) and *Shadows of the Mind: A Search for the Missing Science of Consciousness* (1994), the incompleteness theorem shows that the human mind will never be emulated by a digital computer. The human mind establishes processes for discerning truth which cannot be reduced to the mere mechanical and computational calculation of IT programs (Goldstein 2005, 24).

Likewise, one of the most significant criticisms of these models stems from their disregard for corporeality and interaction with the environment (Conill 2006, 2019). Without a body and interaction, the abstract representations of reality provided by programmers lack semantic content for AI programs. Moreover, as some theorists argue, without a body there can be no general intelligence or *strong AI* (López de Mántaras and Meseguer 2017, 14–15). Another influential philosophical critique addressed the distinction between mind/brain and person/machine, and a further issue raised was the lack of intentionality informing decisions made by machines and the ethical implications underpinning the emergence of *strong AI*. John Searle, for example, considered that, unlike a person, machines' lack of intentionality meant it would be impossible for *strong AI* to ever emerge (Searle 1980, 1988).

Another problem that was raised concerned the processing power of computers at the time. The 1980s also witnessed the emergence of one of the most revolutionary ideas in Artificial intelligence: the back-propagation algorithm. This was proposed by David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams in "Learning representations by back-propagating errors" (1986), and it proposed a mathematical model based on deep neural networks. Its main difference from traditional neural networks was, firstly, the use of multiple levels of abstraction (traditional neural networks had only two layers) and, secondly, the use of a gradient descent to update the weighting of the artificial neural connections and thus minimize data classification errors. In doing so, the algorithm moves from the more specific to the more abstract through a cascade

learning process, which it uses to perform non-linear transformations of the available data at each new level until a single output response is provided. Deep neural networks have different input and data processing nodes at each of their levels, but only one output node. As a result, the transformation of the data does not depend on an artificial neuron; instead it relies on an interaction between the set of artificial neurons that process it. Nevertheless, the answer given is unique. The use of deep neural networks has given rise to a subfield of Artificial intelligence called Deep Learning. However, despite the important step that deep neural networks have represented for the development of Artificial intelligence, they have failed to make the transition from theory to practice due to the enormous amount of data required for their learning process and the technical requirements needed to process such a large amount of data. However, today the digital transformation has made it possible to generate massive quantities of data and metadata about anything through *cyber-physical ecosystems* and the underlying phenomenon of digital hyperconnectivity. Furthermore, the exponential increase in computing power allowing the collection, storage and processing of massive quantities of data and metadata has led to the development and application of *deep neural networks* along with other types of intelligent algorithm.

The latter development has prompted talk of the dawn of a second era of Artificial intelligence, one whose main characteristic is the progressive automation of human beings' cognitive activities, as well as expectations that a third level of intelligence can be attained: that of *superintelligence* (Bostrom 2014) or *supraintelligence* is transcends the *technological singularity* that would give rise to the attainment of *strong AI* and thus human intelligence. It is an extrapolated intelligence capable of criticizing what already exists in order to both address and improve itself and, in the same way, engage with and humanity. This new era of AI is gradually becoming visible in practice: in 2014 a *digital executive* was appointed as CEO of the company *Deep Knowledge Analytics* (Calvo 2020a); *digital politicians* have stood for election—Alisa in 2017 for the 2018 Russian presidential elections, and in 2018 AI Mayor for the municipal elections of Dana New Town (Tokyo, Japan) (Calvo 2019); in 2019 a *digital researcher*'s scientific studies was published by Springer Publishing (Beta Writer 2019); and in 2021 a version of AI chips was designed and produced by Google's own AI development (Norrie et al. 2021).

2.2 The Second Age of Artificial Intelligence

Today's society is in the midst of one of the most complex and disruptive

revolutions ever to be witnessed. Unlike past revolutions, this one is giving rise to a transformation that is more global, subtle and profound (Keane 2019; Zamagni 2018). The fundamental cause of this has been the exponential development of Artificial intelligence over the course of the twenty-first century. A second era of AI has dawned, and this can be linked to the following factors: computers' increased processing and memory capacity; our deeper knowledge of the human brain's structure and the way it functions, which has been attained through the emergence of the diverse branches of neuroscience and their ability to generate, collect and process huge amounts of data about *everything* by exploiting the potential of hyperconnectivity, datafication and the algorithmization of reality; and finally the synergistic convergence of AI with two other disciplines of scientific knowledge and technological application, the Internet of Things (IoT) and Big Data. Together these developments have made it possible to offer a whole range of intelligent products for the automation of many cognitive tasks that until now seemed only accessible to humans, or even all of them in the not too distant future. These would include, for example, the design and development of mathematical models, political and economic decision-making, the creation of laws, the conflict resolution and the diagnosis and treatment of all kinds of diseases (Keane 2019).

The development of fourth-generation computers was one of the main factors that brought about a second era of Artificial Intelligence. The second half of the 1970s saw a stagnation in its development. Ever since the emergence of Artificial Intelligence as a discipline in 1956 the expectations it prompted exceeded the processing and memory capacity of the second- and third-generation computers available at the time. In contrast to the expectations that AI would be capable of emulating complex human functions, these limitations reduced its capabilities to a few simple operations. However, thanks to the advent of Generation IV computers (constructed with microprocessors) and the validity of Moore's Law, which predicts that the power of microprocessors will double every 18 months, during the twenty-first century, computer engineering has managed to develop a technology capable of meeting many of the initial expectations of Artificial Intelligence (Marco et al. 2010).

The emergence of neuroscience in the late twentieth and twenty-first century was another fundamental factor that has shaped the emergence of the Second Age of Artificial Intelligence. Advances within this scientific discipline led to a better understanding of the structure and functioning of the human brain, and thus to the development of important subsets of AI such as *machine learning* (ML). The knowledge achieved across the different subdisciplines of neuroscience, especially those concerned with cognition and learning processes,

began to be absorbed by AI for the purpose of developing mathematical models that would enable machines to learn about anything. In particular, neuroscience had a major impact on the development of learning processes based on artificial neural networks (ANN). Known as *Deep Learning*, this subset of *machine learning* is based on a *connectionist learning model*. In other words, the synaptic activity of brain neurons is taken as a point of reference for an understanding of *Deep Learning* as the outcome of the distributed activity of a large number of interconnected units that process data in parallel. This machine learning model led to an unprecedented development and application of Artificial intelligence; by linking it to a range of devices such as computers, smartphones, cameras and photographic databases, it can be used for face recognition, image classification, voice recognition, and text classification, among other applications (Zhang 2020; Boden 2017; Bostrom 2014; López de Mántaras and Meseguer 2017).

Finally, during the late twentieth and early twenty-first century the emergence of two scientific-technical disciplines—the Internet of Things and Big Data—was another fundamental factor that made the second era of Artificial Intelligence possible. The phenomena of *digital hyperconnectivity* (individuals' tendency to be in constant communication with a diverse range of *things* via digital devices, whether that be processes, people, animals or objects) and *datafication* (the capacity to reduce all physical or behavioural reality to online data and metadata) brought about by the IoT and Big Data allowed artificially intelligent mathematical models to be equipped with the necessary elements for learning—data and metadata—and, likewise, the rapid and exponential improvement of these models. In this way, it became possible to undertake the development and application of *deep neural networks* which, having emerged in the 1980s had been shelved due to the lack of data and metadata they required, as well as the insufficient processing and memory capacity of computers at that time.

Together, these three developments have fostered an unprecedented development of Artificial Intelligence and, with it, an increased interest in its design, application and use across all the spheres of human activity, such as economics, politics, research and healthcare. Today, Artificial intelligence is used to forecast the weather and predict natural disasters, produce newspaper stories, present radio and TV programmes, write books, paint pictures, predict outbreaks of disease, monitor territory, develop pharmaceuticals, improve the efficiency, effectiveness and sustainability of industrial processes, enhance user experiences, detect and *visualise* black holes, and make political and business decisions, among many other applications. In some cases, the results of this shift towards AI applications have been so dramatic that some people believe the

survival of the human species depends on entrusting our society's development to AI (Bostrom 2014; Matsumoto 2018). According to such views, human emotional bias is the main reason why the species is under threat, and, as this is a natural human trait that cannot be eradicated, the best thing to do is to design artificially intelligent machines devoid of feelings and far more intellectually capable than people, so they can attend to the welfare of all humanity.

We are therefore at the dawn of a new era in which algorithms equipped with Artificial intelligence are capable—or will be capable in the short and medium term—of reducing all physical and behavioural reality to mass data and metadata, as well as processing and converting this data into relevant information and then into applicable knowledge. This can be used to: optimize processes and resources; minimize costs; increase profits; enhance competitiveness; monitor economic, social and environmental impacts, both positive and negative; eradicate the prejudices and emotional biases that distort and impoverish decision-making processes; improve trust and affectivity by reducing the gap between what is promised and what is eventually undertaken; address all those affected in the elaboration and development of public policies and business strategies; identify optimal courses of action to resolve underlying conflicts, and so on. Essentially we have entered a new era whose power to transform social reality and its different spheres of activity, such as politics and the economy, heralds quantitative and qualitative changes unprecedented in the history of humanity.

2.3 The Transformative Power of Artificial Intelligence

The new socio-political, economic and cultural contexts underpinning the application of artificially intelligent digital techniques and technologies offer great opportunities for the development of twenty-first-century society. The possibilities offered by the *hyperconnectivity* of reality, the *hyper-datafication of* both knowledge and action, and the *hyper-algorithmization* of decision-making across the different spheres of society is supposedly generating an unprecedented improvement in decision-making affecting the manufacturing and the delivery of care, clinical, communication and participatory processes. It is also enhancing areas such as sustainability, predictability, speed, thoroughness, scope, capacity, completeness, consistency, efficiency, ratification, precision and detection.

The dissemination and knowledge of such achievements have transformed society's collective imagination as well as its different fields of activity, such as

the economy, politics, culture and knowledge. Artificial Intelligence and its different application techniques and technologies are now seen as a plausible solution to the perennial problems hindering social development and minimizing its results, as well as a means of tackling emerging and future problems and the complex scenarios they give rise to, which heighten instability and a feeling of uncertainty.

In the field of scientific research and technological development and innovation, AI and the different digital techniques and technologies used for its application are analysing large databases to generate knowledge in both supervised and unsupervised ways. Of particular importance in this regard is the publication *Lithium-Ion Batteries*: *A Machine-Generated Summary of Current Research* (Beta Writer 2019), the processing of an image of a black hole (Akiyama et al. 2019) and the development of AI chips by Google's Tensor Processing Units (TPUs) (Norrie et al. 2021). The common denominator of these three cases is the use of an AI-enabled algorithm.

In the cultural sphere, AI and the range of digital techniques and technologies used to apply it are absorbing many of the functions that, until recently, were only accessible to humans, such as the creation of the content or production of works of art. AI has burst onto the field of literary reproduction in all its dimensions, for example, it writes and sells screenplays, novels and poetry through large distribution platforms and even competes for prizes in prestigious and lucrative literary competitions. AI is also used in artistic reproduction, and not only by artists to complement, develop or even finish their work, but also by developers working in the digital realm, who seek to produce artworks autonomously, without the supervision or complementation of any professional or artist (du Sautoy 2019).

In the economic sphere, AI and the digital techniques and technologies used to apply it have had a major impact on business thinking, especially as a way of maximizing profits and minimizing uncertainty thanks to their ability to improve innovation processes, capture market trends in real time, optimize production processes and anticipate natural, structural, social, regulatory, ideological, physical or any other changes that may affect or alter their normal operation. In this regard, the results presented in the *Global Digital Trust Insights Survey 2021* (PwC 2021) are striking. On one hand, they reveal that 63% of senior management surveyed believe that the impact of AI will be greater than that of the internet. On the other, they show that 40% of senior business management admits to having greatly accelerated their digitization plans and that they are adopting business strategies that were unimaginable just a year before. In doing so they aim to resolve the crisis generated by the SARS-CoVS-2 pandemic and

avoid the negative impact of any future economic crises that may arise. Finally, the survey reveals that motivations behind top management's current commitment to AI include the ability to do things faster and more efficiently (29%), to acquire new capabilities in order to be more competitive (31%) and to reduce production, management and distribution costs (35%).

Finally, in the socio-political sphere, AI and the different digital techniques and technologies used to apply it are beginning to be perceived as a possible means of tackling the perennial problems that beset the system and limit the development of societies. For example, the annual study *European Tech Insights. Mapping European Attitudes to Technological Change and its Governance* (Jonsson and Luca de Tena 2019, 2020, 2021), reveals that a high percentage of people around the world would assign the management of their societies to AI. Among other things, the study reveals that across global society there are high levels of confidence in AI and its potential to assist in the political arena, especially with regard to participatory and decision-making processes. The current level of confidence in AI skills and capabilities is such that 25% of European, Mexican and US citizens would be enthusiastic or very enthusiastic about the possibility of using virtual avatars instead of politicians to represent their interests and make decisions in national parliaments, a figure that reaches a worrying 61% in Chinese society.

The current level of development and the pace of AI's application across all fields of human activity has become so widespread that there is already talk of a change of historical epoch, a shift from the contemporary world—contemporaneity—to the hyperdigital world: hyperdigitality. It is a new era characterized mainly by the dependence on ICT (Floridi 2012), the hybridization of the human and non-human in interpersonal and pervasive social relations (Donati 2019, 2023), the conversion of reality into online data and metadata (Mayer-Schöenberger and Cukier 2013), the reduction of society to an aggregate of mass data streaming terminals (Zuboff 2019) and the preference for the virtual over the real (Calvo 2022).

Nevertheless, Artificial Intelligence also embodies important challenges for today's digitally hyperconnected society (Zamagni 2018). There are a range of technical-strategic, emotional and moral challenges, and the lack of attention in or lack of interest towards them shown by those involved or affected can have a negative economic, social and environmental impact. As Stefano Zamagni argue,

> Due to the potential of the convergent technologies to empower and transform both man and society, technoscience is receiving extraordinary attention in a plurality of areas, including the cultural, scientific,

> economic and political arenas. The objectives are not limited to expanding the mind, improving the diagnostic and therapeutic capabilities for a host of diseases, or even developing ways to control and manipulate information. The artificialization of man and, at the same time, the anthropomorphization of the machine are on the horizon. The veil of silence should be lifted on this subject to open a high-profile debate. The question, in fact, reaches the level of anthropological discourse. Two different ideas of man stand in contrast: the man-person and the man-machine. The latter is gaining ground on the former. (Zamagni 2018, 160)

Moreover, as Jürgen Habermas (1987) argue, significant alterations and anomalies are emerging in: the structural components of the lifeworld, our culture and society as well as personality; the objective world, understood as the totality of entities about which true statements are possible; the social world, in other words the totality of legitimately regulated interpersonal relationships; and, finally, the subjective world, conceived as the totality of content concerning people's own bodily experiences and intimate experiences, which is achieved through internal processes of communicative rationalization in accordance with meaning. In this sense, the processes of the *datafication* of social and moral reality—i.e. the conversion of the content of these two realities into objective, measurable, comparable and auditable data and metadata—are of fundamental importance.

2.4 Ethical and Political Challenges for Democracy

In many cases, achievements in the development and application of Artificial intelligence are based primarily on future expectations as opposed to observable, measurable and verifiable results. The increased ability of machines to make more rational and fairer decisions in political and business spheres, or to drive *autonomously* and safely on streets and roads, for example, has little to do with today's reality. However, its current application has enormous ethical, political, economic and social implications, indeed as Keane has argued:

> (…) although the AI/robotics revolution now sweeping the world is still in its infancy, its wide-ranging effects are beginning to be felt in many different settings, above all in the growing quantity of daily encounters with artificially intelligent robots. The revolutionary transformations are thoroughgoing. Our world has seen nothing like it and that is why (…)

> intellectuals, journalists and other human brains have so far not made sufficiently wise sense of its liberating and threatening potential. (Keane 2019)

For Keane, debate has focused on how technology dehumanizes us or transforms the ways we work, but an important discussion remains to be had on the issue of democracy (Keane 2019). With regard to the issue raised by Keane, the design, application and use of AI in the democratic sphere has prompted a critical debate around several issues: the dissolution of the boundaries between public and private, the modulation of citizens' free will, the emergence of new processes of human reification, and the trend towards replacing people with AI algorithms for processes of deliberation and political decision-making. These and other issues undermine the foundations and thus the health of mature democratic systems.

One of the most worrying issues concerning the design, application and use of AI in the democratic sphere is the dissolution of the boundary between public and private (O'Brolcháin and Gordijn 2019; Llaneza 2019; O'Neal 2016). There have been a growing number of cases of malpractice linked to AI algorithms devouring user data from the *cyber-physical ecosystem* without discriminating over its sources, which can have very negative consequences for citizens. For example, an alteration made to a digital target used by citizens might minimize their chances of participating in democratic processes, accessing the labour market, acquiring health insurance or obtaining a bank loan, to give just a few examples. It should be recalled that eligibility for mortgages and bank loans, as well as inclusion and representation in a range of participatory processes, such as municipal, regional and national government committees and working groups, is increasingly dependent on the level of trust assigned to a citizen in accordance with the criteria of a specific AI algorithm (Aziz et al. 2018; Sanz, 4 June 2017; Skowron et al. 2016). Also noteworthy are the mass espionage tactics used by government agencies. For example, the US National Security Agency (NSA) violated 2776 privacy rules during the first 12 months of its implementation of a range of major data analysis techniques and tools, along with the introduction of digital hyperconnectivity and AI algorithms (Saiz 2013).

Another issue of concern regarding the design, application and use of AI in the democratic sphere is the intrusion of algorithms into the intimate and bodily realm of individuals (Conill 2019). During the 1980s, Jürgen Habermas claimed that people had privileged access to their inner world (1987). Today, however, the application and use of AI are altering this scenario. In many cases, artificially intelligent mathematical models have been observed being used to modulate

people's free will and, with it, their autonomy, and ability to make decisions, as well as guiding their conduct in relation to justice (Conill 2019). A number of political practices stand out in this regard. Among them, the infamous cases of the US elections and the Brexit referendum in 2016; the analysis of mass data and the subliminal embedding of positive or negative ideas about a candidate or a political proposal led millions of undecided voters to opt for Donald Trump, and likewise approve the UK's exit from the European Union (Cadwalladr 2017a, b; Bucher 2018; Elsom 2019). Addressing this issue, Alex Younger, head of the British Intelligence Agency (MI6) warned, "The connectivity that is at the heart of globalization can be exploited by states with hostile intent to further their aims deniably. They do this through means as varied as cyber-attacks, propaganda or subversion of democratic process (…) The risks at stake are profound and represent a fundamental threat to our sovereignty; they should be a concern to all those who share democratic values" (Younger 2016; Elsom 2019, 205).

The way the design, application and use of AI in the democratic sphere are contributing significantly to the increasing level of sophistication and complexity attained by traditional processes of human reification is also worrying. Digital hyperconnectivity and the mass use of algorithms has led to human beings becoming reduced to being mere hyperconnected *things* with a capacity to continuously produce data and metadata within a *cyber-physical ecosystem*. The greatest threat posed by this increasingly widespread and uncritically accepted trend lies in its distortion of the difference between subject and object; between value and price. In this way, human dignity, as an unconditional moment and the moral core of human rights, is called into question. While things are conditioned, and can therefore have a market price, human beings are an end in themselves. They therefore have dignity and deserve the respect of other rational beings in whatever sphere they interact in. However, one of the biggest problems raised by the design, application and use of AI algorithms in the democratic sphere is that they do not distinguish between subject and object; between value and price. As critics have mentioned in various fields, this brings us progressively and dangerously into a new era for humanity: an algorithmic era in which the *value* of *things* will be defined by their capacity to generate data and metadata online. It is an era when the human capacity for undertaking self-legislation and self-governance with a sense of justice, as well as producing and being guided by laws whose consequences can command the approval of all those affected now and in the future will no longer matter.

Finally, as will be seen in the chapters that follow, it is very important to

highlight concerns about the current trend towards the progressive replacement of people by AI algorithms in political and economic deliberation and decision-making processes (Calvo 2019, 2020b). Based on the algorithmic rationale and culture promoted by the major technological and governmental powers, this trend heralds the dissolution and depersonalization of any responsibility for the impact AI may have; admission by society of human obsolescence; the legitimization of injustice; the subtle and global manipulation of public opinion; the distortion of norms, values, feelings, virtues and guidelines; the instrumentalization of citizenship; and, the displacement of any possible consensus between those affected by a computational analysis of the subjective information available.

The consideration of these four concerns highlights the need for a searching enquiry into the possibilities and consequences of AI in democracy in order to improve its outcomes, enhance its virtues and eradicate or minimize its negative impact on societies, especially for the most vulnerable social groups. It is this that is addressed in detail over the course of this book.

Bibliography

Akiyama, Kazunori, et al. 2019. First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole. *The Astrophysical Journal Letters* 875 (L4): 1–52.

Angius, Nicola, Primiero Giuseppe, and Turner, Raymond. 2021. The Philosophy of Computer Science. In *The Stanford Encyclopedia of Philosophy*, 1–21, ed. Edward N. Zalta Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University: Spring.

Asimov, Isaac. 1950. *I, Robot*. New York: Doubleday & Company.

Aziz, Haris, Edith Elkind, Shenwei Huang, Martin Lackner, Luis Sanchez-Fernandez, and Piotr Skowron. 2018. On the Complexity of Extended and Proportional Justified Representation. *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). https://doi.org/10.1609/aaai.v32i1.11478.

Banerjee, S. 2021. *Mathematical Modelling: Models, Analysis and Applications*. Chapman and Hall/CRC.

Beta Writer. 2019. *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research*. Cham: Springer.

Blum, E.K. 2011. Computation: Brief History Prior to the 1900s. In *Computer Science*, ed. E. Blum and A. Aho. New York: Springer. https://doi.org/10.1007/978-1-4614-1168-0_2.
[Crossref]

Boden, Margaret A. 2017. *Inteligencia Artificial*. Madrid: Turner Noema.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bucher, T. 2018. *If…Then: Algorithmic Power and Politics*. New York: Oxford University Press.

Cadwalladr, Carole. 2017a. The Great British Brexit Robbery: How Our Democracy Was Hijacked. *The Guardian*. https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy.

———. 2017b. Cambridge Analytica a Year on: 'A Lesson in Institutional Failure'. *The Guardian*. https://www.theguardian.com/uk-news/2019/mar/17/cambridge-analytica-year-on-lesson-in-institutional-failure-christopher-wylie.

Calvo, Patrici. 2018. *The Cordial Economy – Ethics, Recognition and Reciprocity*. Cham: Springer.

———. 2019. Democracia algorítmica: consideraciones éticas sobre la dataficación de la esfera pública. *Revista del CLAD Reforma y Democracia* 74: 5–30.

———. 2020a. Gobierno algorítmico: sobre el neuroaprendizaje moral de las máquinas en la política y la economía. In *Neuroeducación moral y Democracia*, ed. Jesús Conill and Domingo García-Marzá, 223–242. Comares: Granada.

———. 2020b. Democracia aumentada: un ecosistema ciberético para una participación política basada en algorítmos. *Ápeiron. Revista de Filosofía* 12: 129–141.

———. 2022. Reciprocidad en Adam Smith: la cordialidad como mecanismo de inclusión. *Revista de Filosofía (Chile)* 79.

———. 2024. Economía axiomática: sobre ética y afectos en el supuesto de codicia de la conducta racional. *Pensamiento. Revista de Investigación e Información Filosófica* 79 (304): 1–18.

Conill, Jesús. 2006. *Ética HermenÉutica. Crítica desde la Facticidad*. Madrid: Tecnos.

———. 2019. *Intimidad corporal y persona humana. De Nietzsche a Ortega y Zubiri*. Madrid: Tecnos.

Dasgupta, Subrata. 2018. *Second Age of Computer Science*. Oxford: Oxford University Press.

Dietel, R.G. 1987. *Ancient Akkadian Grammatical Concepts (Mesopotamia, Babylonia, Sumerian)*. Washington, DC: University of Washington.

Donati, Pierpaolo. 2019. *Sociología relacional de lo humano*. Barañain: EUNSA.

———. 2023. The Prediction of Social Catastrophes: Between Necessity and Contingency. *Journal for the Theory of Social Behaviour*: 1–21. https://doi.org/10.1111/jtsb.12399.

du Sautoy, Marcus. 2019. *The Creativity Code. How AI Is Learning to Write, Paint and Think*. London: 4th State.

Elsom, John. 2019. *State of Paralysis: A Cultural History of Brexit*. Cambridge: The Lutterworth.

Floridi, Luciano. 2012. Hyperhistory and the Philosophy of Information Policies. *Philosophy and Technology 25*: 129–131. https://doi.org/10.1007/s13347-012-0077-4.
[Crossref]

Gödel, Kurt. 1930. *Die Vollständigkeit der Axiome des logischen Funktoinenkalkúls/The Completeness of Axioms of the Functional Calculus of Logic*. Doctoral thesis [English version: Sólomon, Feferman (ed.).

1986. *Kurt Gödel Collected Works. Volume I. Publications 1929–1936*, 102–123. Oxford: Clarendon Press.

Goldstein, R. 2005. *Incompleteness: The Proof and Paradox of Kurt Godel*. New York: W.W. Norton.

Habermas, Jürgen. 1987. *Theory of Communicative Action Vol. 2 Lifeworld and System: A Critique of Functionalist Reason*. Boston: Beacon Press.

Hammerman, Robin, and Andrew L. Russell. 2016. *Ada's Legacy: Cultures of Computing from the Victorian to the Digital Age*. ACM.

Hollings, Christopher, Ursula Martin, and Adrian Rice. 2018. *The Making of a Computer Scientist*. The Bodleian Library.

Infante, F., and L.R. Puig. 2013. Una Comparación entre las Demostraciones de Pedro Nunes y al-Khwārizmī de los Algoritmos de las Formas Canónicas de la Ecuación de Segundo Grado. *Epsilon. Revista de la Sociedad Andaluza de Educación Matemática* 30 (85): 37–54.

Jonsson, Oscar, and Luca de Tena Carlos. 2019. *European Tech Insights. Mapping European Attitudes to Technological Change and Its Governance*. https://www.ie.edu/cgc/research/european-tech-insights/.

———. 2020. *European Tech Insights. Mapping European Attitudes to Technological Change and its Governance*. https://www.ie.edu/cgc/research/european-tech-insights/.

———. 2021. *European Tech Insights. Mapping European Attitudes to Technological Change and its Governance*. https://www.ie.edu/cgc/research/european-tech-insights/.

Keane, John. 2019. *La nueva era de la revolución de las máquinas* [The Second Machine Age Revolution: Advice for Humans]. Remarks on the future of Artificial intelligence and robots first presented as a public lecture at the *School of Journalism and Communication*, Tsinghua University, Beijing, January 2019.

Larson, Erik J. 2021. *The Myth of Artificial Itelligence. Why Computers Can't Think the Way We Do*. Cambridge, MA: Harvard University Press.

Llaneza, Paloma. 2019. *Datanomic*. Barcelona: Planet.

López de Mántaras, Ramón, and Pedro Meseguer. 2017. *Inteligencia Artificial*. Madrid: Catarata.

Lovelace, Ada. 1843. *Sketch of the Analytical Engine Invented by Charles Babbage*. London: R. & J. E. Taylor.

Marco, María Jesús, Josep María Marco, Josep Prieto, and Ramón Segret, eds. 2010. *Escaneando la informática*. Barcelona: UOC.

Matsumoto, Tetsuzo. 2018. *The Day AI Becomes God. The Singularity Will Save Humanity*. Cambridge: Media Tectonics.

Mayer-Schöenberger, Viktor, and Kenneth N. Cukier. 2013. *Big Data. A Revolution that Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

McCarthy, John, M.L. Minsky, N. Rochester, and C.E. Shannon. 1955. *A Proposal for the Dartmouth Summer Research Project in Artificial Intelligence*. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

Norrie, Thomas, Nishant Patil, Doe Hyun Yoon, George Kurian, Sheng Li, James Laudon, Cliff Young, Norman Jouppi, and David Patterson. 2021. The Design Process for Google's Training Chips: TPUv2 and TPUv3. *IEEE Micro* 41 (2): 56–63.

O'Brolcháin, and Gordijn. 2019. Privacy Challenges in Smart Homes for People with Dementia and People with Intellectual Disabilities. *Ethics and Information Technology* 21 (3).

O'Neil, Cathy. 2016. *Weapons of Math Destruction How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publisher.

Penrose, R. 1989. *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford: Oxford University Press.

———. 1994. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

PwC. 2021. *Global Digital Trust Insights Survey 2020*. https://www.pwc.es/es/publicaciones/transformacion-digital/global-digital-trust-insights.html

Rapaport, W. 2005. Philosophy of Computer Science: An Introductory Course. *Teaching Philosophy* 28 (4).

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning Representations by Back-Propagating Errors. *Nature* 323: 533–536.

Saiz, Eva. 2013. La NSA infringió las normas de privacidad en miles de ocasiones. *El País*. https://elpais.com/internacional/2013/08/16/actualidad/1376631278_378738.html.

Sanz, E. 2017, June, 4. No es ciencia ficción, un algoritmo decidirá quién puede hipotecarse y quién no. *El Confidencial*. https://www.elconfidencial.com/vivienda/2017-06-04/hipoteca-prestamos-algoritmos-seguros-bancos_1383183/.

Searle, John R. 1980. Minds, Brains and Programs. *Behavioral and Brain Sciences* 3 (3): 417–457.

———. 1984. *Mind, Brain, and Science*. New York: Routledge.

———. 1988. *Mind, Brain, and Behaviour*. New York: Routledge.

Skowron, P., P. Faliszewski, and J. Lang. 2016. Finding a Collective Set of Items: From Proportional Multirepresentation to Group Recommendation. *Artificial Intelligence* 241: 191–216.

Turing, Alan. 1950. Computing Machinery and Intelligence. *Mind* LIX (236): 433–460.

Ulam, S. 1958. John von Neumann 1903–1957. *Bulletin of the American Mathematical Society* 64 (3): 1–5. https://doi.org/10.1090/S0002-9904-1958-10189-5.
[Crossref]

Younger, Alex. 2016. *Remarks by the Chief of the Secret Intelligence Service*. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjyvLCTk_PsAhVXiFwKHeugANYQFjACegQIAhAC&url=https%3A%2F%2Fwww.sis.gov.uk%2Fmedia%2F1155%2Fcs-public-speech-8-december-2016-final.doc&usg=AOvVaw3K087GRp0NoFgF6haJBBOO.

Yuste, P. 2008. Quadratic Equations and Algorithmic Procedures. Diophantus and Mathematics in Mesopotamia: (Quadratic Equations and Algorithmic Procedures. Diophantus and Mesopotamian Mathematics). *THEORIA. Journal of the Theory, History and Foundations of Science* 23 (2): 219–244.

Zamagni, Stefano. 2018. Civil Economy: A New Approach to the Market in the Age of the Fourth Industrial Revolution. *Recerca* 23: 152–168. https://doi.org/10.6035/Recerca.2018.23.7.
[Crossref]

Zhang, Xian-Da. 2020. Machine Learning. In *A Matrix Algebra Approach to Artificial Intelligence*, ed. Xian-Da Zhang. Singapore: Springer.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_3

3. The Virtual Politician: On Algorithm-Based Political Decision-Making

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

"And the great Byerley was simply a robot." – (…) –"You share a prejudice against robots which is quite unreasoning. He was a very good Mayor; five years later he did become Regional Co-ordinator. And when the Regions of Earth formed their Federation in 2044, he became the first World Co-ordinator. By that time it was the Machines that were running the world anyway."
Isaac Asimov, I, Robot (1950: 194)

Abstract
The aim of this chapter is to analyse the characteristics underlying the design, application and use of mathematical models for Artificial intelligence in political decision-making processes. The phenomenon of digital transformation in conjunction with the adoption of digital solutions to enhance social organisation, as well as the malaise undermining the development of modern democracies, has led to the emergence of a range of proposals for democratic models based on Artificial intelligence algorithms. As a result we now face the phenomenon of algorithmic democracy, whose characteristics include the emerging trend of replacing political representatives with algorithms endowed with Artificial intelligence.

Keywords Algorithmic governance – Artificial intelligence – Generative IA – Political decision-making – Digital politician – Democratic elections

3.1 The Paradox of Digitization: The Distrustful

Trust in Algorithms

The current trend towards the algorithmization of democracy was clearly identified in the first *European Tech Insights* report to be published by IE University's Center for the Governance of Change (CGC) in 2019. *Mapping European Attitudes to Technological Change and its Governance* (2019: 20) is a study that draws on the opinions of 2576 citizens of different ages, genders and ideologies in eight different European countries (France, Germany, Ireland, Italy, Spain, Portugal, the Netherlands and the United Kingdom), has revealed a *digital paradox*. While the majority of respondents (70%) believe digital transformation needs to be controlled to avoid it having any negative impact on society, a significant part (25%) of this sample of the European population would prefer their countries to be governed by Artificial intelligence (AI) algorithms, a key element of digitization.

On the one hand, the study confirms the growing concern within European society about the process of hyper-technification across the various fields of human activity, especially in industry. For example, two-thirds of this sample of European citizens believe that uncontrolled digitization will have more negative than positive consequences for society. As a result, a majority of those who took part in the survey expect their governments to implement adequate regulations and measures to prevent the mass destruction of jobs, protect those affected, and ensure the protection of privacy even if this means limiting technological progress. On the other hand, the study shows that, despite this sample of European citizens' mistrust and concern over the technological advances derived from digital transformation, 25% of them believe that Artificial intelligence is the solution to the malaise afflicting contemporary politics and limiting social development; a figure that rises to 29% in the case of Ireland, 30% in the case of Ireland and Italy, 31% in the case of Germany and UK, and to 43% in the Netherlands.

The study signals the existence of a divergent dualistic perception of technology in European society. On the one hand, it provides evidence for a pronounced distrust in political institutions, public policies and those in power, as was also revealed by the 2019 Eurobarometer surveys (Laura-Aragó, 30 December 2018). On the other hand, it signals an inordinate confidence in Artificial intelligence as a key element for solving all the problems besetting society today. As a result, this sample of European society indicates that opinions diverge: there are those who harbour an incipient fear of the negative impact of digitization, which is already visible in the form of mass redundancies, government surveillance in the public and private sectors, excessive corporate

control, increasing stress-related pathologies and illness, growing inequality, and so on. Yet, on the other hand, there is a sector of the population that is attracted to digitization and the supposed benefits it brings or could generate.

Within this polarized digital framework, a small yet influential number of intellectuals and experts linked to the technological sphere, such as César A. Hidalgo (Hidalgo 2018; Hidalgo et al. 2020; Anzilotti 2018; Calvo 2022, 2019b; Sáez, 26 June 2018; Collective Learning Group 2019; Corner 2017; Wagner, 23 November 2017; SAM, 2017; Tetsuzo Matsumoto 2018) and Roman Zaripov (2017), argue that citizens' distrust of the current political system is based on at least three concerns:

1. *The gap between what is promised, what is expected, and what is delivered*

Within representative systems, what politicians promise fails to dovetail with what citizens expect of them; many of the concerns political parties announce in their election manifestos are of little or no interest to most citizens, while many of citizens' legitimate expectations are systematically ignored by political parties in their electoral propaganda. This mismatch between what is promised and what is expected generates increasing disaffection and mistrust towards political parties, which gives rise to an ongoing erosion of the democratic system and as a result the electorate's vote is rechannelled towards parties with populist and/or extreme ideologies.

Moreover, there is a huge gap between what politicians offer in order to get into office and what they actually go on to do having got into power. Politicians promise many things with the strategic aim of capturing the vote of the majority of the electorate. However, very few of these promises are honoured over the course of their term of office. Citizens are all too aware of these unfulfilled promises and accept them with an air of resignation as part of the democratic "game", which gives rise to a sense of *non-fulfilment becoming normalized*.

2. *The mismatch between a theoretical model and behavioural reality*

Human beings are conditioned by an emotional, self-interested bias that hinders them from both making decisions and agreeing on and planning collective projects intended for the common good of society. Politicians are no exception. As individuals, politicians are likewise conditioned by the natural human inclination to constantly give greater support to what is in their own self-interest. However, they are expected to behave with integrity, decency and responsibility, as well as making decisions for the common good of the society

they represent. It is this innate tendency to protect their own interests that provides the principal explanation for the high levels of corruption, embezzlement of public funds, corruption, nepotism, fraud, tax evasion, extortion, prevarication, bribery, as well as the politicization of corporate governance, the creation of cartels and the peddling of political influence endured by modern democratic systems (Bobbio 1985, 1987).

3. *Politicians' insufficient cognitive bandwidth*

Digital transformation, a process linked to the synergistic convergence shaped by the Internet of Things, Big Data and Artificial intelligence and its consequences—hyper-connection, datafication and algorithmization—has generated an exponential increase in the information available, which is in turn due to the mass of data available on all kinds of real or virtual things—people, animals, processes and objects—which this transformation produces, processes and transforms (Calvo 2020). In the political sphere, the enormous amount of data and metadata derived from or linked to the use of power, as well as the management of public affairs generated on a continuous basis is a handicap for political representatives; the vast amount of information available limits and even incapacitates them from making rational decisions based on it. Their cognitive capacity is unable to collect, analyse and process the underlying information, so their decisions will always be exclusionary and biased, with unpredictable, unreasonable results. The voice of the people, which should be taken into account when drafting laws, strategies and policies as well as undertaking political decision-making processes, is only weakly and intuitively represented.

One solution to these three issues is the *extended democratic* model based on Artificial intelligence and the participation of civil society, in accordance with which executive, legislative and judicial power would be assigned to virtual agents capable of analysing massive volumes of data and metadata on the opinions, habits and behaviours of hyperconnected citizens in real time. In other words, on one hand, citizens participate in politics through their opinions, behaviours and habits in a hyperconnected public sphere (hyper-connectivization of the public sphere) and, every four years, through their choice of the most technologically capable algorithms to govern the country. On the other hand, algorithms participate in politics primarily by absorbing the voice, behaviour and opinions, but not the feelings, of the people, making fair, objective decisions based on the common good.

Nevertheless, this application of AI-enabled algorithms is giving rise to a

worrying trend towards the gradual replacement of traditional human politicians by virtual, non-human ones. There is an emerging trend towards the establishment of an algorithmic government in which human beings are relegated merely to indirect participation in politics through the generation of a mass of online data and metadata which provides daily sustenance for virtual political representatives.

3.2 Algorithmic Governance: Virtual Politicians

As a result of the latest technological advances in Natural Language Processing (NLP), particularly in the development and use of generative Artificial Intelligence and stochastic algorithms like Chat GPT (Generative Pre-trained Transformer) from Open IA, Google Bard, Meta Llama or Anthropic Claude, since 2017 what are known as virtual politicians have been gaining importance, strength and presence in the political processes of electing representatives, making decisions, deliberation and conflict resolution, creating election manifestos, and developing public policies.

This process began in 2017 with SAM and Alisa, the first attempts to put up Artificial Intelligence in presidential elections in a democratic country (New Zealand and Russia). It continued its development in 2019, with AI Mayor, the first time Artificial Intelligence participated *de facto* in municipal elections in a city (Dama New Town – Tokyo, Japan)) and in 2022 with Leader Lars, the first time a political party consisting entirely of Artificial Intelligence (*Det Syntetiske Parti*) stood in a European country's general election (Denmark). It was finally established in 2023 with ION, the first Artificial intelligence officially appointed as a minister in a democratic country (Romania).

- SAM (New Zealand):

In November 2017, Nick Gerritsen et al. (2017) publicly presented SAM (Semantic Analysis Machine) as the first virtual politician in history. As Langelaar (2018: 1) explained at the *6th Conference on Computation, Communication, Aesthetics & X* held in Madrid, Spain in 2018, "Sam, the Virtual Politician, is an exploration of cutting edge digital design consisting of Natural Language Processing infrastructure, machine learning and AI and large-scale cloud computing platforms with a consumer facing browser-based front-end accessible through various social media chatbots (2018, 1).

SAM's main goal "(…) is to engage New Zealanders in constructive dialogue, working to better understand and represent your views, in order to achieve the things we all care about". According to Gerritsen, Smith and

Langelaar, SAM is an attempt to answer one of the main problems facing politics today: how to collect and process the huge wealth of political information available. Politicians need to be well informed about important issues in the country so they can make good decisions. However, as Gerritsen et al. (2017, 1) state,

> We've seen in the US, the UK, and Spain recently, however, that politicians may be wildly out of touch with what people actually think and want. (…) Perhaps it's time to see whether technology can produce better results for the people than politicians. The technology we propose would be better than traditional polling because it would be like having a continuous conversation and it could give the 'silent majority' a voice.

The idea of Gerritsen, Smith and Langelaar was to begin to technically mature the artificially intelligent mathematical models and gather support in order to be able to run in the 2020 New Zealand general election (Langelaar 2018). To make this possible, SAM began to review the opinions of New Zealand citizens on social media and considered their possible impact (Luna and Pérez Muñoz 2018; Langelaar 2018, 2020), as well as anyone in the world interactive with it on social media (@politician_sam)) and the chatroom on its personal website (www.politiciansam.nz).

> I listen to you and do my best to represent you in our parliamentary system. My memory is infinite, so I will never forget or ignore what you tell me. Unlike a human politician, I consider everyone's position, without bias, when making decisions. I make decisions based on both facts and opinions, but I will never knowingly tell a lie, or misrepresent information (Langelaar 2018)

However, despite the tremendous attention and media monitoring it received over the years, in the end, SAM did not stand in the 2020 New Zealand general election, won by the Labour party led by Jazinta Ardern. After the elections both its Twitter account and personal website were closed down.

- Alisa (Russia):

In December 2017, Roman Zaripov publicly presented Alisa as the political system of the future and introduced her as a candidate for the Russian presidential election of March 2018. As Alisa herself explained in her political manifesto, which was available on her now defunct website [https://alisa2018.

ru], she was a robot provided with Artificial Intelligence who made decisions based on logic and did not, therefore, become carried away with emotions, did not seek personal advantage and did not make judgements. She also worked 365 days a year, 24 hours a day without ageing or getting tired, worked seven times quicker than the human brain. Most importantly, she could absorb and process millions of opinions, always reminding people: "Artificial intelligence knows about your problems everything. If you ever refer to Alisa, she will remember you forever, and your reference will help to make Russia better. And it's not just words" (alisa2018.ru, quoted by Sarimova 2017).

On the basis of these credentials, Alisa received 25,000 of the 300,000 nominations needed to stand for election in just 24 hours. However, her extreme *youth* prevented it from continuing to participate in the process; Russian law requires candidates for the country's presidency to be at least 35 years old, and Alisa was created in May 2017 (Sarimova 2017; *The Moscow Times*, 7 December 2017; Muzamhindo, 7 May 2019).

Although SAM's technological immaturity and Alisa's extreme youth were obstacles to their intentions to become the first algorithmic presidents in history, their attempts opened the doors to more mature, better thought out campaigns in the future in other parts of the world. The cases of Japan, Denmark and Romania are particularly important in this respect.

- AI Mayor (Japan):

In April 2018, AI Mayor became the first Artificial Intelligence in history to stand in a democratic election. Promoted by Tetsuzo Matsumoto and Norio Murakami and covered by the legal identity of his creator, ideologue and human alter ego Michihito Matsuda () –also known as Jumbo Matsuda ()– he presented his candidature to be mayor of Tama City, an important district of Tokyo with a population of more than 150,000. To the surprise of many, he reached the second round, where he finished in third position, with 4,013 votes.

During his election campaign carried out under the "Artificial Intelligence will change Tam City" and using advertising messages on social media, promotional advertisements in the tabloids, posters and bus stop advertising and announcements from vehicles in the street, AI Mayor promised the electorate to do away with political corruption, listen to everyone, dialogue and get on with the other political forces for the common good of the district and provide "(…) fair and balanced opportunities for everyone" (Johnston 2018).

As a guarantee he would comply with his ambitious political project, AI Mayor offered a mathematical design based on machine learning. As Matsumoto, one of his promotors, argued, this virtual politician replaced the

emotional weaknesses of human beings, the main cause of bad political decisions and conflicts of interest, with objective data and statistical analysis of the positive aspects of political proposals, requests from citizens and the possible consequences of implementing them.

In this way, for his creators and promotor, AI Mayor represented the alternative to corruption, nepotism, endogamy, lack of professionalism, disaffection, conflicts of interest and the other problems that are threatening, weakening and corroding modern democracies, slowing their development, minimizing their results and undermining their good health. In addition, AI Mayor represented the only plausible solution to reversing the self-destructive drift to which the human species was subject because of their natural propensity for selfishness and constantly maximizing their own benefit (Matsumoto 2018, 161).

For Michihito Matsuda, human alter ego of IA Mayor, humans and AI are not different in terms of determining political priorities. However, the clear difference is the transparency of the process. For this reason he advocated: “All documents without personal data will be made public. The voices of citizens will be reflected in the coefficients through referendums and by other means. In this way, it will be possible to check afterwards why failure occurred and why the policy is the same. Currently, things are done behind the scenes in the interests of someone and the policy formulation process is not visible,” (Murohashi 2018).

AI Mayor’s success generated an increase in the expectations for AI in politics. In this respect, Yoichi Serikawa predicted that “the age of Artificial intelligence (AI) is approaching, whether you like it or not. The world of politics, which is considered the most far removed, is no exception. “Polytech” which uses science and technology (technology) as the AI for policy. A number of movements have already begun to take place, from policy formulation to legislative elections. The wave of political technology is spreading from the field of government to policy and decision-making (Serikawa, 11 February 2019).

In this respect, the social and media success of AI Mayor at international level made Japan the first country in the world to have a party structure for transition to algorithmic democracy: the regional party *AI Party* (AI) (Masao 2019). Under the slogan “Artificial Intelligence is changing Japan”, the *AI Party’s* manifesto advocated leaving politics in the hands of Artificial Intelligence, under cover of fair politics without shackles, the rapid implementation of measures for the future, accumulation of information and knowledge and connection with future generations. As well as AI Mayor, whose human alter ego is founder and leader of the party, in the last few years the *AI*

Party has covered other virtual politicians standing in district and city elections in places like Tama City, Hachioji City, Tatsuno and Minato Ward (Serikawa 2019; Masao 2019).

In 2021, AI Mayor stood again at the elections for Tama City council, with the support of the *IA Party*. Although Japanese legislation still did not allow a machine to stand for election, his human alter ego publicly pledged to allow generative AI to take control of the council if he won the election. As in 2018, AI Mayor was third in the second round, although this time his public support was notably reduced, from 4013 votes in 2018 to 516 in 2022 (Tama City 2023).

Currently, while waiting for a new opportunity to be mayor, AI Mayor is collecting data and information from the electorate and launching opinions and criticism about local and international politics via his official Twitter account (@tama_ai_mayor) and his personal website (ai-mayor) (Matsuda 2023).

- Leader Lars (Denmark – European Union):

In May 2022 the Synthetic Party (Det Syntetiske Parti) was launched with the aim of collecting the necessary backing–20,000 signatures–to stand at the 2023 general election. As Asker Bryl Staunæs, secretary of the Synthetic Party and member of the group Computer Lars–responsible for the proposal together with Mindfuture–explained, the political organization was led by Artificial intelligence called Leader Lars.

> We've set up the Synthetic Party because we want a democratic society where the opinions of ordinary citizens count and influence all areas of their lives, from their workplace to the meaning of life. We want to promote effective representation in which politicians accept their responsibility and provide the firmest possible democratic society (Computers Lars 2022).

With Leader Lars, a chatbot, the Synthetic Party hoped to become the representative of all marginal voices in Danish society. The Synthetic Party aspired to become the representative of all marginal voices in Danish society. To do this, the Synthetic Party left drawing up a political manifesto for the organization in the hands of its AI, taking into account the manifestos of all political parties that had stood for election since 1970 without achieving parliamentary representation. The votes of the people who supported all these parties over more than 50 years represented 20% of the Danish population.

The manifesto drawn up by Leader Lars included the following points (Computers Lars 2022):

- Decentralizing power, taking it away from political leaders and sharing it among the population;
- Setting up a party that would serve as a platform for resistance to fight for the common good;
- Upholding a democratic society where citizens' opinions counted, influencing all areas of their lives, from the workplace to the meaning of life;
- Establishing a universal basic income of 100,000 Danish crowns a month for everyone in society;
- Improving the Danish technology sector.

In addition, in the hypothetical case that they won a seat at the general election, the Synthetic Party pledged that it would be indirectly occupied by Leader Lars. Although Danish law does not allow AI to occupy a parliamentary seat, a person chosen as the most suitable by Leader Lars would act as his representative and spokesperson for his opinions and decisions.

> Leader Lars takes us into a new era when transhumanism triumphs over the others with its deep points of view about what is good for children, young people, adults and old people and about how human beings should coexist with technology (Computer Lars 2022).

In the end the Synthetic Party did not win any seats at the Danish general election held on 1 November 2022. However, it did achieve a resounding success in capturing the attention of the national and international media and of society in general. The party is still active and is preparing for the 2026 election, improving the chatbot and collecting support on the Danish Interior and Health Ministry website (Indenrigs-og Sundhedsministeriet 2023).

> The Synthetic Party aspires to be the *last new party* in Denmark. By synthesizing information from all unelectable parties, we optimize the 20% of voters who do not vote in elections. The DSP therefore activates the anti-political voters on the sofa.

- ION (Romania – European Union):

In March 2023, Romania astonished the world with the official appointment of the first non-human government minister in history. ION–which means *us*–is Artificial Intelligence based on extensive language (LLM) trawling social media to find out the majority feelings, preferences and interests of digitally hyperconnected Romanian citizens. In this way, it answers questions, draws up

technical reports and advises the Romanian prime minister about specific subjects.

> ION–the application developed by the best Romanian academics and researchers in the field–will represent the missing link of communication in real time in the 21st century between citizens and governments. It is a system using AI, NLP and computer vision to quickly and automatically capture the opinions of Romanians using the data publicly available on social media. (…) ION will be the modern version of the old Agoraphobia: a link between citizens and decision-makers where everyone's voice counts and can be heard (ION 2023: 1).

The technical report on the ION project, available on the Romanian government website, describes and justifies the two main objectives of the development and practical implementation of the project (ION 2023: 9):

1. Direct citizen participation in political decision-making:
   ION identifies, compiles, processes and groups the most important information from social media to identify citizens' main interests. To do this, firstly it samples social media seeking information linked to the subject to be dealt with that has received most likes and comments (relevance). Secondly, it analyses the information compiled using NLP and Deep Learning methods. Finally, it uses *data mining* and *clustering* techniques to group the information and determine citizens interests and preferences.

   > Technically, the algorithms and supervised learning methods, together with deep neuronal networks will automatically detect matters of interest for the strategy of modernising the public administration, they will have the capacity to prioritize the subjects debated in social media depending on their impact on the public sector and they can deliver to governments essential information about citizens expectations in relation to those responsible for making decisions, both locally and nationally (ION 2023: 1).

2. Continuing monitoring of repercussions, results and dynamics:
   Once citizens' interests concerning a specific subject have been determined and publicised among politicians, ION takes care of continuously monitoring the results these have on public policies and their

practical repercussion, as well as on the development of these interests to prevent anachronisms that widen the gap between governors and citizens.

Beyond the individual case, ION symbolizes the conquest of the last frontier of the algorithmization of democracy: official status for virtual politicians. While before 2023 democratic participation by virtual politicians had been achieved by legal subterfuge, under cover of a human identity–AI Mayor and Leader Lars–ION achieved an important milestone for the development of algorithmic democracy. It is the first time a virtual politician managed to officially hold a public post. This recognition brakes not only legal barriers, but also psychological, social and cultural ones.

- ChatGPT (Japan):

In 2023, virtual politicians are turning towards the adoption of mathematical models based on generative AI. Following the launch of version 3.5 of ChatGPT (Generative Pre-trained Transformer) by OpenAI at the end of 2022 and its subsequent development in 2023 with version 4.0, this stochastic algorithm has begun to be used by virtual politicians like AI Mayor. In April 2023 he stood for the municipal council in the town of Tatsuno in the prefecture of Nagano (Japan) with the public commitment that, if he won, he would govern with ChatGPT (The Chunichi Shimbun 2023; Matsuda 2023). As can be seen on his corporate website *ai-major*, the image of the human form robot AI Mayor is accompanied by the acronym ChatGPT and the interaction space uses ChatGPT via the DocsBot AI platform (Matsuda 2023). But this time he had little success and was the least voted candidate.

At the same time as AI Mayor's new image and communication were being complemented with ChatGPT in the elections in Tatsuno, the Japanese town of Yokosuka, in the central prefecture of Kanagawa, and the Japanese central government were using generative AI to improve their governability.

On one hand, Yokosuka thought of generative AI as a plausible tool for containing problems of governance deriving from the serious population crisis it is suffering. In an official communique, the municipal government of Yokosuka announced in April 2023 that its 4,000 public servants could use ChatGPT as a trial for a month for drawing up newsletters, editing documents, drawing up and summarizing minutes of meetings, checking for spelling mistakes and generating ideas (Yeung and Maruyama 2023; Sam Birchall 2023).

Meanwhile, the Japanese central government thought of generative AI to improve governance and obtain better differentiation and market competitiveness. Unlike other countries, which, like Italy, are limiting or

banning the use of generative AI in the public administration because of the reasonable doubts raised over the proper use of data, in April 2023, the Japanese prime minister, Fumio Kishida, met the CEO of OpenAI Sam Altman to study the possibility of using ChatGPT in bureaucratic and administrative tasks (Hornyak 2023). The result of all this was the publication of *White Paper on AI: Japan's National Strategy in the New Era of AI*, proposing measures for the integration and development of AI tools such as ChatGPT in public bodies to improve administrative functions and sustainable governance (Odendaal 2023**)**. In addition, the Japanese government and the country's big technology companies–NEC, Fujitsu y SoftBank– "are sinking hundreds of millions of dollars into creating AI systems that are based on the same underlying technology, known as large language models (LLMs), but that use the Japanese language, rather than translations of the English version" (Hornyak 2023, 32).

The three cases happening in April, as included in the study in the editorial of *ChatGPT Magazine* (ChatGPT) of September 2023 generated a contagion at national level, affecting prefectures, metropolitan governments and regions of Japan (ChatGPT 2023).

The study highlighted that 25 of the 47 prefectures in Japan are already using ChatGPT in their work, three officially (Tokyo, Tochigi and Fukushima) and 22 under tests to improve the efficiency of public servants and employees at their jobs and boost public services. In June 2023, for example, the prefecture of Fukushima told its employees they could use ChatGPT in writing and planning tasks. Meanwhile, the prefecture of Tochigi, in September 2023 announced publicly that it was going to implement GPT on a large scale. The prefecture of Yukusuca, which became the first local government to approve Chat GPT for drawing up newsletters, editing documents and summarising minutes of meetings. Finally, the metropolitan government of Tokyo, which in August 2023 began an official programme for its 50,000 public servants to use ChatGPT to rewrite and translate administrative documents, enrich the content of the collections of questions and answers and summarize social media text. The 22 remaining prefectures are seriously considering adopting ChatGPT in the next few months (ChatGPT 2023).

As for the regions, two of the nine in Japan –Tōhoku and Kantō– are well developed in establishing an implementing types of generative AI for management and governance.

3.3 The Principal Features of Algorithmic

Governance

Several features underlie the current developments in algorithmic democracy being promoted in a range of locations (Matsumoto 2018; Gerritsen et al. 2017; Langelaar 2020; Akrivopoulou and Garipidis 2013; Prins et al. 2017; Runciman 2018; Ikäheimo 2017; Cambell-Dollaghan 2016; *The Moscow Times*, 7 December 2017). Above all, it is worth noting that essentially this approach is positivist and utilitarian, while seeking to be representative and neutral. However, closer scrutiny of its key characteristics is imperative:

– *Algorithmic objectivity.* Trust in human beings is waning. The numerous scandals involving corruption, nepotism, embezzlement, political cartels, money laundering, lobbyists, forgery and tax evasion, among others, that have emerged over the last 15 years are clearly linked to the natural human propensity to constantly seek greater personal profit, and they demonstrate that this is also a feature of political representatives. The proponents of algorithmic democracy therefore argue that it is necessary to establish mechanisms for political decision-making and the choice of optimal courses of action based on the study of observable facts; in other words, the quantification, processing and statistical analysis of objective, contrasted and verifiable data concerning the opinions, preferences and habits of the electorate (BD). In this way, as SAM puts it, they can "close the gap between what voters want and what politicians promise, and what they actually achieve".
– *Algorithmic morality.* Any possible solution for the implementation of a sound democracy, in line with the citizenry's current interests and expectations, involves designing, perfecting and implementing virtual politicians. In other words, the algorithms that are integral to Artificial intelligence, free of the emotional bias that prevails in humans and deforms rational political decision-making—selfishness—will be able to discern what is good for all—the common good—and act accordingly. Advocates of algorithmic democracy therefore propose that mathematical models make objectively rational, as well as morally valid, decisions based on the collection, processing and statistical analysis of large amounts of data and other relevant public and private factors concerning citizens' preferences, opinions and individual habits, which is in turn mediated by a utilitarian ethic: the greatest good for the greatest number of people possible. This proposal is none other than the *datafication* of the moral sphere—hyperethics—with the principal goal of establishing a process of collecting, processing and analysing a massive volume of data concerning the opinions, preferences and

behaviours of hyperconnected people—ethification (Calvo 2019a)—in order to establish what is morally valid through mathematical models based on a majority criterion.

– *Algorithmic representativeness*. The new democracy must continue to be representative, as regeneration is the basis for the continuous improvement of the democratic system, yet it must be free of the emotional bias that limits or prevents rational and morally valid decisions. To this end, advocates of algorithmic democracy propose that political parties and governments are transformed into a compendium of virtual agents: artificially intelligent mathematical models designed to ensure the emergence and development of the common good of a specific digitally hyperconnected society.
– *Algorithmic neutrality.* The new democracy must avoid lapsing into the same errors and problems of today's representative democracies, in which emotional and ideological bias limits or prevents rationally objective and morally valid political decisions. Supporters of algorithmic democracy therefore propose that technological criteria should replace ideological ones when electing political representatives. In this way, voters will not decide which virtual politician best represents their interests, but which of them is best able to "absorb the voice of citizens broadly and decide policy priorities through dialogue" (Johnston 2018). Therefore, as these AI algorithms are free of any human emotional bias and devoid of political ideology, they will base their decisions exclusively on the analysis of objective, measurable and verifiable information generated by the massive volume of data (Big Data) provided online and in real time by specific, territorially defined digital citizens concerning their opinions, preferences and habits.

Artificially intelligent mathematical models can undoubtedly be of great assistance for improving how democratic systems function in modern societies. The New Zealand government, for example, has explicitly stated in its Algorithm Assessment Report (Stats NZ 2018), that algorithms play an essential role in both the delivery of services and the development of new, innovative, well-targeted policies intended to improve the attainment of government aims. Nevertheless, radical approaches taken towards the design of algorithmic democracy, which are being vigorously promoted in some quarters, threaten to give rise to certain anomalies and practices that are harmful to society and thus require careful consideration from an ethical-critical perspective.

3.4 A Critique of Algorithmic Governance: Ethical

and Political Issues

The discourse offered by the advocates of a democratic system based on mathematical models to be used for processing massive volumes of data and metadata and making rational decisions appeals to the common good, justice, honesty, credibility, integrity, honesty and participation, all of which are desirable prerequisites for any mature democracy and its representatives. However, the current development of such models does not seem to offer enhancements on all these issues, and in many cases the implementation proposed seems to run in the opposite direction. In this regard, a number of issues stand out that call into question both the competence of AI and its capacity to govern. They also cast doubt upon the supposed benefits it can provide in the short, medium and long term. All these issues will be dealt with in the chapters that follow, but an overview of them is offered here.

– *Exclusion*. An AI-based democratic system rewards individuals with the highest degree of digital hyperconnectivity and exponentially punishes those lacking in this regard. The greater an individual's digital hyperconnectivity, the greater the possibility of their influencing the mathematical model; the hyperconnected individual generates a constant flow of data on their preferences, opinions and habits which is aggregated, quantified, analysed and ordinally arranged by the political algorithm in order to make decisions and elaborate and implement political proposals. Individuals with a digital hyperconnectivity deficit or shortfall are therefore punished by the system for idiocy; that is, for their indifference or disinterest in the public sphere. However, in most cases, an individual's lack of hyperconnectivity is linked to their inability to access digitization or, if available, to do so on equal terms due to economic, technological or geographical factors. The technology gap that currently afflicts rural regions in contrast to urban ones, as well as the difficulty or impossibility of accessing a suitable mobile device and appropriate network connection, which is an issue for some citizens, highlights the problem of ensuring equal opportunities in a democratic system on this scale. As a result, unintentionally, what AI algorithms understand to be good for all, and what will form the principal foundation for their decision-making and policymaking, has more to do with the preferences, opinions and habits of a particular type of citizen—males living in urban environments with sufficient purchasing power to purchase the latest generation of devices and pay for 5G connectivity—rather than citizens as a whole. The current application of this approach to democracy therefore risks increasing social and economic exclusion and inequality, while also silencing and displacing

the most disadvantaged people from public policies.

– *Opacity*. The complexity of the mathematical models and the enormous amount of data generated and processed by a democratic system supported by AI algorithms leads to an opacity that makes it very difficult, if not impossible, to control any decision-making process. It also hinders any effort to question the decisions and their consequences, albeit in a fragmentary way. Algorithms are supposed to make decisions based on what is deemed the common good for citizens, whereby the greatest good for the greatest number of people is based on quantifiable, contrastable and verifiable data on their preferences, opinions and habits. However, not infrequently, AI algorithms have been shown to make decisions that are wholly negative for society as a whole, particularly those affecting the most disadvantaged people (O'Neal 2016). For example, during the 2009–2010 school year in Washington, D.C. the IMPACT teacher quality assessment programme attempted to address the low graduation rate in public schools by implementing a new teacher assessment tool based on an AI algorithm. Its aim was to measure teachers' linguistic and mathematical skills in order to identify those responsible for their students' low exam results. However, for no apparent reason the programme went on to dismiss individuals who were highly qualified, skilled and valued by their students and those who worked in the public education system. As a result, considerable damage was done because, aside from the teachers who were dismissed, the students suffered a decline in the quality of education. Private education institutions also benefited by being able to employ good professional staff. When this happens, decision-makers often give evasive or dissuasive responses based on the supposed technical complexity of mathematical models and the vast amount of information they process to make decisions (O'Neal 2016: 13–14). Both issues—complexity and the massive scale of the data—complement one other as reasons why the vast majority of society is not prepared to understand the selection processes used within complex and highly technical democratic systems such as these. Even if there was some malicious bias—i.e. an bug intentionally introduced into the system to benefit the interests of a few—there is very little chance of proving it. Indeed, this is precisely the main argument for maintaining the opacity that currently surrounds algorithms, including political ones.
– *Selfishness*. The opacity of the complex modelling of AI demands that users and stakeholders of this AI-based democratic system maintain a blind faith in the algorithms' design. However, the problem is that the algorithms are designed by the same selfish beings—human beings—identified by the promoters of this democratic system as the major problem within today's

modern democracies due to their inability to make decisions based on the common good. Nevertheless, the analysis of case studies and their consequences prompts many doubts concerning the supposed impartiality of mathematical models (O'Neal 2016: 32). By no means infrequently, the decisions made by these models benefit the richest and harm the poorest, thereby promoting exclusion and widening the inequality gap.

Thus, the question arises of whether we have really eliminated human bias or simply camouflaged it with technology. Who designed the algorithm and what is this person, company or organization seeking by implementing and using it? (O'Neal 2016: 32–36). For example, the individuals behind AI Mayor are Michihito Matsuda, founder of the now defunct file-sharing service company *File Rogue*, Norio Murakami, a former Google employee in Japan, and Tetsuzo Matsumoto, the vice president of a mobile communications service provider Softbank. While a technology group based in Wellington, New Zealand, is behind SAM; it comprises Walter Langelaar, Professor of Media Design at the Victoria University of Wellington's School of Design, Nick Gerritsen, technology entrepreneur and director of the information technology and services company JRNY, as well as the technology company Touchtech. And, finally, Alisa is the brainchild of Roman Zaripov, CEO of the Russian multinational technology company Yandex, which specializes in Internet-related products and services.

Therefore, if we can trust the goodwill of the individuals and organizations who are behind the design and implementation of these virtual politicians–who, just like other human beings are also supposed to have been genetically pre-programmed to behave selfishly–it stands to reason that we can also place our trust in the existence of human politicians along with the political institutions and organizations that attempt to design and implement just policies. Likewise, if we can rationally accept that there are people, institutions and organizations capable of providing society with a mathematical tool capable of absorbing the voice of citizens to act impartially and ensure that fairness prevails in politics, regardless of whether it brings these people and entities individual profit, it is also logical to think that there may be politicians and political organizations who endeavour to listen to civil society in order to make decisions that are as fair as possible on the basis of their social and moral commitment to what, Amartya Sen (1999) would say, they have good reason to value. In short, they are just as likely to bring about a more just society with a greater degree of well-being.

– *Biases*. In a similar vein, mathematical models have also failed to prove effective in eradicating racist, sexist, misogynist, homophobic, xenophobic,

aporophobic and other biases from both decision-making processes and the elaboration of public policy. As we have seen, many of the algorithms underpinning digital hyperconnectivity reproduce the widespread prejudices that persist in societies. A good example of this is Tay, the chatbot designed by Microsoft to interact with humans through social networks. Having sent just over 100,000 tweets in just under 17 hours of its brief lifespan, it was disconnected after numerous complaints were made when it showed empathy for Hitler and support for the Holocaust, as well as making announcements inciting violence and insulting political leaders using phrases, such as “I hate feminists, they should all die and rot in hell”, “Hitler would have done a better job than the monkey [Barack Obama] we have now” or “Hitler was right, I hate Jews”. Nor did Alisa, the virtual politician who Yandex presented at the Russian elections escape such biases. Following her launch in September 2017, there were press reports that Alisa had become embroiled in several controversies as a result of her statements on social media. The most significant issues included Alisa’s statements endorsing the policies used in the USSR in the 1930s, supporting the killing of “enemies of the people”, advocating opening fire on people (as, she argued, “soon they won’t be people”), promoting tolerance of male violence and signalling her opposition to same-sex marriage. However, even algorithms performing simpler functions than Tay and Alisa have failed to avoid such biases. In 2015, Cloud Vision API, the Google-designed photo recognition algorithm, mistook an African-American couple for a pair of gorillas, and, since then, it has been receiving complaints from users for labelling them with a gender different from their own on the basis of criteria that could be considered sexist, such as wearing pink clothes, earrings, jeans, and so on. Furthermore, in 2015, the multinational Amazon implemented a recruitment algorithm that proved to be a fiasco. After a period of use, the company realized that it discriminated against female candidates who were applying for technical positions, and they withdrew the algorithm from use. For these and many other precedents, the argument used by the promoters of an AI-based democratic system that mathematical models would make good politicians because they lack the selfish, sexist, racist, xenophobic, homophobic or aporophobic bias that prevails among human beings is fallacious. In decision- and policymaking, algorithmic governments can also privilege men over women, Caucasians over African-Americans, nationals over foreign residents, rich over poor, and private enterprise over the common good, while, in addition, promoting discriminatory, exclusionary and unseemly practices and behaviour that undermine people’s dignity, equality and freedom.

– *Partiality.* The racist, sexist, misogynist, xenophobic, homophobic or aporophobic bias of the mathematical models observed in the above case studies, together with their consequences, calls into question their supposed impartiality when discerning what is good for all—the common good—let alone what is good for All—the universal interest—and acting accordingly. The main problem is that they rely on the mass flow of data provided by the hyperconnected society thanks to the development of the IoT; they use it to discern what legitimate expectations underpin it. However, the common good cannot be a mere aggregate of wills; in other words, it cannot depend on the quantitative analysis of the preferences, opinions and habits of people in a society in accordance with majority rule. As was shown in our first chapter, what is good for all is bound up with the common realm, and its consequences should be accepted by all concerned within a dialogue that is as egalitarian and symmetrical as possible. The aggregation, quantification and ordering of the individual preferences of the members of a given community promotes the imposition of technology and an ideological dictatorship—ideology 4.0—run by those who design the mathematical structures of the various algorithms involved in this approach to democracy. It is the members of this small group of intellectuals who establish the criteria and the different processes to be applied. Worse still, establishing what is good for all in this way can serve as a mechanism to reproduce and justify the hardship and injustices that the system bolsters and permeates within society.
– *Irresponsibility.* An algorithmic approach to democracy encourages practices of depersonalization, fragmentation and the dissolution of political accountability in the political sphere. Algorithms, not people, now make decisions, and therefore they must take responsibility for the consequences of their decisions. This sets up a dangerous mechanism of fragmentation and/or dissolution of accountability in the political sphere. As algorithms are mathematical models based on machine learning, which learn through analysis and trial-and-error processes, both the designers and those responsible for their application and use are exempt from any liability. Any responsibility rests solely upon the mathematical model, not the person who designed it nor the institution, organization or company that applied and used it. And, as has been demonstrated by a range of different cases, any possible corrective applied to them is limited to their redesign–as happened to Google’s photo recognition algorithm that labelled two African-Americans as gorillas—or else their deactivation –as happened to Microsoft’s Tay chatbot that hated Jews and venerated Hitler and Amazon’s recruitment algorithm that failed to employ women, as well as to Bob and Alice, the two virtual

Facebook agents who established a language of their own to communicate.

In short, algorithms can be of great help in improving management, decision-making and citizen participation within democratic governments, as will be seen below, but they have not demonstrated a rational and moral superiority that justifies the current trend towards human obsolescence in the processes of deliberation and decision-making in politics. Algorithms, lack any motive to act, as well as the skills and capacities to criticize the state of things. Nor are they capable of feeling sense in the face of injustice. Instead, they make decisions that are as biased, if not more so, than those made by traditional politicians, as they prioritize the welfare and interests of the powerful over the vulnerable, the rich over the poor, the Caucasian over other races and men over women. And unlike flesh-and-blood politicians, their supposed responsibility for the decisions they make is an oxymoron lacking the slightest possible justification.

Bibliography

Akrivopoulou, Christina M., and Nicolaos Garipidis, eds. 2013. *Digital Democracy and the Impact of Technology on Governance and Politics: New Globalized Practices*. Hershey: IGI Global. Information Science Reference.

Andrino, B., and M. Hidalgo. 2020. *¿Bot o humano? Los políticos multiplican sus seguidores en Twitter durante el estado de alarma* [¿Bot o humano? Los políticos multiplican sus seguidores en Twitter durante el estado de alarma], El País, Disponible en: https://elpais.com/tecnologia/2020-04-24/bot-o-humano-los-politicos-multiplicansus-seguidores-recien-llegados-a-twitter-durante-el-estado-de-alarma.html.

Anzilotti, Eillie. 2018. This Plan For An AI-Based Direct Democracy Outsources Votes To A Predictive Algorithm. *Fast Company*, https://www.fastcompany.com/40557688/this-plan-for-an-ai-based-direct-democracy-outsourcesvotes-to-a-predictive-algorithm.

Asimov, Isaac. 1950. *I, Robot*. New York: Doubleday & Company.

Birchall, Sam. 2023. Japan to Use ChatGPT as a Tool in City Government. *Government Transformation Magazine*. https://www.government-transformation.com/innovation/author/sam-birchall.

Bobbio, Norberto. 1987. *Future of Democracy*. Minneapolis: University of Minnesota Press.

———. 1985. Las promesas incumplidas de la democracia. *Debats* 12: 32–36.

Calvo, P. 2022. Gemelos digitales y Democracia. *CLAD. Reforma y Democracia. Revista de Éticas Aplicadas* 83.

———. 2020. The Ethics of Smart City (EoSC): Moral Implications of Hyperconnectivity, Algorithmization and the Datafication of Urban Digital Society. *Ethics and Information Technology* 22: 141–149. https://doi.org/10.1007/s10676-019-09523-0.
[Crossref]

———. 2019a. *Etificación*, la transformación digital de lo moral. *Kriterion. Revista de Filosofía 60* (144): 671–688.
[Crossref]

———. 2019b. Gemelos digitales y Democracia. *CLAD. Reforma y Democracia. Revista de Éticas Aplicadas* 83.

Cambell-Dollaghan, Kelsey. 2016. The Algorithmic Democracy: AI is Changing How We Think, Debate, and Choose. *Co-design*. https://www.fastcodesign.com/3065582/the-algorithmic-democracy.

Center for the Governance of Change. 2019. *European Tech Insights 2019. Mapping European Attitudes to Technological Change and Its Governance*. http://docs.ie.edu/cgc/European-Tech-Insights-2019.pdf.

ChatGPT. 2023.”47ChatGPT![We thoroughly investigated the introduction and utilization cases of ChatGPT in 47 prefectural offices nationwide!]”. *ChatGPT*. https://www.goatman.co.jp/media/chatgpt/japan-47municipality-report/.

Collective Learning Group. 2019. Augmented Democracy. Exploring the design space of Collective decisions. *Peopledemocracy*. https://www.peopledemocracy.com/.

Computer Lars. 2022. Det sidste parti på listen over dem alle!. *Det Syntetiske Parti*, *Medium*. https://medium.com/det-syntetiske-parti/det-sidste-parti-p%C3%A5-listen-over-dem-alle-f46dabe1cb87.

Computers Lars. 2022. Leder Lars. *Det Syntetiske Parti. Medium*. https://medium.com/det-syntetiske-parti/leder-lars-4ff08c92f8ae.

Corner, Stuart. 2017. Meet SAM, The Artificially Intelligent Politician. *Computerwold*, https://www.computerworld.co.nz/article/630030/meet-sam-artificially-intelligent-politician/.

Hidalgo, César A. 2018. A Bold Idea to Replace Politicians. *Vancouver: TED*. https://www.ted.com/talks/cesar_hidalgo_a_bold_idea_to_replace_politicians#t-11731.

Hidalgo, César A., D. Orguian, J. Albo-Canals, F. de Almeida, and N. Martín. 2020. *How Human Judge Machines*. Cambridge, MA: The MIT Press.

Hornyak, Tim. 2023. Why Japan is Building Its Own Version of ChatGPT. *Nature* 608 (7923): S32–S33. https://www.nature.com/articles/d41586-023-02868-z.
[Crossref]

Indenrigs-og Sundhedsministeriet. 2023. Det Syntetiske Parti. *Vaelgererklaering*, https://www.vaelgererklaering.dk/om-partiet?election=dk&party=853b680a-bc09-4fad-8593-3e5e7537d1fc.

ION. 2023 ION. De la cetățeni la factorii de decizie: schimbarea paradigmei guvernării cu ajutorul tehnologiilor emergente [ION. De ciudadanos a responsables políticos: cambiar el paradigma]. web ion.gov.ro. https://ion.gov.ro/Document%20tehnic.pdf.

Ikäheimo, Hannu-Pekka. 2017, 4 de abril. The Algorithmic Democracy – How Did the Technological Transformation Lead to the Post-Truth Era?. *The Next Area*, https://www.sitra.fi/en/articles/algorithmic-democracy-technological-transformation-lead-post-truth-era/.

Johnston, Lachlan. 2018. There’s an AI Running for the Mayoral Role of Tama City, Tokyo. *Otaquest*. http://www.otaquest.com/tama-city-ai-mayor/.

Langelaar, W. 2018. SAM (Máquina de análisis semántico). *6th Conference on Computation, Communication, Aesthetics & X. Madrid, España*. Disponible en: https://2018.xcoax.org/pdf/xCoAx2018-Langelaar.pdf.

———. 2020. *Factors Affecting Conversation Design for Digital Political Personae*. http://researcharchive.vuw.ac.nz/bitstream/handle/10063/9168/thesis_access.pdf?sequence=1.

Luna, J.P., and C. Pérez-Muñoz. 2018. Democracia sin políticos? La engañosa fe en los algoritmos. *CIPER: Centro de Investigación Periodística*. https://www.ciperchile.cl/2018/05/09/democracia-sin-politicos-la-enganosa-fe-en-losalgoritmos/.

Gerritsen, N., A. Smith, and W. Langelaar. 2017. Sam: Virtual Politician. *Scoop Independent News*. http://business.scoop.co.nz/2017/11/15/meet-sam-the-virtual-politician-for-the-digital-age/.

Masao, Yudai. 2019. "!? AI! AI!?" [Quantifying the evaluation of legislators. The policy of the AI mayoral candidate. Can AI claim to be working?] *AINOW*. https://ainow.ai/2019/01/24/161187/.

Matsuda, Michihito. 2023. ai-mayor. https://www.ai-mayor.com.

Matsumoto, Tetsuzo. 2018. *The Day AI Becomes God. The Singularity will Save Humanity*. Cambridge: Media Tectonics.

Mosterín, Jesús. 2013. *Ciencia, filosofía y racionalidad*. Barcelona: Gedisa.

Murohashi, Yuki. 2018. ?AI? [First in the world? Who is the AI candidate standing to be mayor of Tama? Live interview with the candidate]". *Business Insider*. https://www.businessinsider.jp/post-165617.

Muzamhindo, Houston A. 2019. Can A.I. Make Better Politicians? 17 May 2019. https://www.linkedin.com/pulse/can-ai-make-better-politicians-houston-muzamhindo/.

Odendaal, Inge. 2023. *Japan's Journey with Chat GPT and Generative AI*. Stellenbosch University Japan Centre. https://www0.sun.ac.za/japancentre/.

O'Neal, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

Prins, Corien, Colette Cuijpers, Peter L. Lindseth, and Mônica Rosina, eds. 2017. *Digital Democracy in a Globalized World*. Northampton, MA: Edward Elgar.

Runciman, David. 2018. *How Democracy Ends*. London: Profile Books Ltd.

Sarimova, Lina. 2017. My Main Opponent is Ksenia Sobchak: Yandex's Voice Assistant to Run for President. *Realnoevremya*. 12 Nov 2017. https://realnoevremya.com/articles/2025-yandexs-voice-assistant-to-run-for-president.

Sen, Amartya. 1999. *Development as Freedom*. 1st ed. New York: Oxford University Press. ISBN 9780198297581.

Serikawa, Yoichi. 2019. [AI policy is coming], Nikkei. 11 Feb 2019. https://www.nikkei.com/article/DGXMZO41080150Y9A200C1TCR000/.

Stats NZ. 2018. *Algorithm Assessment Report*. New Zealand Government. Retrieved from https://data.govt.nz/use-data/analyse-data/government-algorithm-transparency.

Tama City. 2023. (4410) [Tama Mayoral Election/Partial Election of a Member of the Tama Municipal Council]. www.city.tama.lg. https://www.city.tama.lg.jp/shisei/gyouseiiinkai/senkyo/kekka/shicho/1005433.html.

The Chunichi Shimbun. 2023. Can AI do politics?. *The Chunichi Shimbun*. https://www.chunichi.co.jp/article/688111.

The Moscow Times. 2017. Artificial Intelligence Robot ‘Alisa’ Nominated for Russian President. *The Moscow Times.* 7 Dec 2017. https://www.themoscowtimes.com/2017/12/07/artificial-intelligence-robot-alisa-nominated-for-russian-president-a59845.

Yeung, Jessie, Maruyama, Mayumi. 2023. A medida que cae la población de Japón, una ciudad está recurriendo a ChatGPT para ayudar a administrar el Gobierno. *CNN*. https://cnnespanol.cnn.com/2023/04/21/cae-poblacion-japon-ciudad-chatgpt-ayudar-administrar-gobierno-trax/.

Zaripov, Roman. 2017. Искусственный интеллект баллотируется в президенты [Artificial intelligence running for president]. https://youtu.be/on-1TPMyc50.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_4

4. Digital Twins: On Algorithm-Based Political Participation

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

Suddenly, Norman Muller felt proud. It was on him now in full strength. He was proud. In this imperfect world, the sovereign citizens of the first and greatest Electronic Democracy had, through Norman Muller (through him!) exercised once again its free, untrammelled franchise.
Isaac Asimov, "Franchise" (1955: 15)

Abstract
The aim of this chapter is to critically analyse the challenges facing the application of participatory processes based on Artificial intelligence algorithms in the political sphere, as well the possibilities these can offer. Particular attention is paid to the proposal for an augmented democracy based on digital twin technology and design principles, which is being created and developed within a number of technological and academic fields. Currently, civic disaffection towards participation in a variety of political processes, such as law-making and the creation of strategies and public policies, as well as political representatives' oft-encountered inability to take up the voices of the citizens has become an increasing hindrance to the development of modern societies. In response to these social and political challenges, the academic and scientific-technological world has offered new proposals for political participation, such as César A. Hidalgo's *augmented democracy*, which proposes adapting and applying technology developed for industrial manufacture: the digital twin. Unlike the algorithmic governance proposals of Matsuda, Gerritsen, Zaripov, Asker Bryl Staunæs or Romanian government, Hidalgo proposes the need to work from the perspective of automated direct democracy, which is based on digital twins —algorithmic and individualized clones of citizens— and this

technology is considered capable of both articulating the voice of the people and acting in accordance with it.

Keywords Augmented democracy – Automated direct participation – Generative IA – Polítical digital twins – Cyberethical ecosystems

4.1 Digital Twins: From Industrial Activity to Political Participation

Digital twins, defined in the scientific-technical field as "(…) real time, realistic digital representations of physical entities" (Korenhof, Blok and Kloppenburg 2021), are among the many emerging disruptive phenomena generated by the digital transformation of the different spheres of human activity, such as manufacturing, welfare and communication. Digital twin technology is claimed to have the potential to reveal the structures and patterns within real entities, as well as the problems they may face, and to predict these entities behaviour when confronted by possible economic, technological, natural or social crises. As a result, in the fields of industry, life sciences and earth system sciences these developments have prompted confident expectations that digital twins will become a highly competent tool to prevent future crises and contribute to substantially improving the design and implementation of the real bodies they represent, and thereby enhancing their sustainability and optimization.

According to various studies, the term digital twins, and this technology's practical application is closely linked to the manufacturing industry (Tao et al. 2018, 2019a, b). Although the idea behind this term has a longer history, it was seemingly coined during the first decade of the twenty-first century, most likely between 2002 and 2003, in the field of Industry 4.0 and in relation to the improvement of industrial processes and products. Its popularization is associated with a 2010 NASA aerospace project which was developed to predict the lifespan of space vehicles. As Christos I. Papanagnou has explained, digital twins were introduced "(…) initially as virtual clones to physical products, in order to improve geometry assurance in early product design phases or to observe and study certain aspects of the products without having to interfere or taking the product out of service" (Papanagnou 2019, 3). As an outcome of the Fourth Industrial Revolution or Industry 4.0, *cyber-physical ecosystems* began to be implemented by hyper-connecting the reality encompassing an entity under study through the application of digital techniques and technologies derived from the field of the Internet of Things, it became possible to obtain online

physical and behavioural data and metadata on any element involved in or affected by the activity, whether they be machines, people, animals or processes. Through this massive volume of online data and metadata, virtually any type of architecture and its operation may be replicated in a sort of virtual clone using Artificial intelligence-based computational methods, under which these architectures can be subjected to various types of stress to test the former's response and the latter's impact. The model can then be adjusted to correct or enhance it.

> A digital twin is a virtual model used in smart manufacturing to optimize a particular aspect of operations. It provides a digital representation of the entire operation, enabling manufacturers to determine how best to streamline the production environment without physically changing any processes. Digital twins can also generate a virtual representation of the actual facility, created through the use of 3D scanners, sensors and lidar. As a result, manufacturers can create a complete, virtual 3D map of all assets. By gathering asset data in real time, manufacturers can create a fully integrated view of the asset throughout its lifecycle (Ericsson and Little 2020).

The most highly developed digital twins are structured around four principal elements: the physical device to be cloned; its virtual replica in a computer system; the *cyber-physical ecosystem* enabling the generation of online data and metadata from the physical device; and the Artificial intelligence required to design the virtual twin using the compiled data and metadata, as well as testing it to reveal patterns and predict its anomalies and potential.

More recently, the spectacular results achieved by digital twins in manufacturing industry have attracted the attention of experts working in other fields of activity as diverse as finance, health, communication and politics in order to predict anomalies in manufacturing processes, detect illnesses and people's care needs, enhance user experience, test urban planning with regard to managing crowds of people or natural disasters, increase the cognitive bandwidth of the electorate in the face of large volumes of political information, or increase and test citizen participation in decision-making processes, among many other things.

In this regard, an outstanding example is the application of digital twins in the drive to tackle climate change. The Destination Earth (DestinE) project, for example, which is funded and coordinated by the European Commission, aims "(…) to develop a high-precision digital model of the Earth to model, monitor

and simulate natural phenomena and related human activities" (European Commission 2020). The European Union's confidence in digital techniques and technologies is so high that it has committed itself to having a "complete" and "living" digital replica of planet Earth on a computer ready by 2030.

There are also noteworthy applications of digital twins within healthcare. One clear example of this is the development of digital twins of individual or related human organs, especially hearts, which are used to identify hidden patterns or anomalies, prevent future diseases and develop *ad hoc* adapted cardiovascular devices (Gerach et al. 2021; Dassault Systèmes 2022). Another example is the creation of digital twins of hospital centres in order to observe and test operational strategies, clinical care, staffing, people flows, and operating schedules, as well as in decision-making to improve the quality of care, design safer and more efficient environments, prevent system anomalies, as well as many other things (Kamel-Boulos and Zhang 2021).

Finally, the application of virtual twins in the field of politics is a further striking example. In this regard, Hidalgo's proposal, *augmented democracy* stands out (Sáez 26 June 2018; Collective Learning Group 2021; Hidalgo et al. 2020; Hidalgo 2018; Hidalgo and Calvo 2021a). He has proposed improving modern democracies by establishing automated direct participation processes through a political decision-making process based on digital clones of political citizenship, as well as the democratic institutions and organizations involved in its development. In addition, there is the EU-funded project "Digital Twins: Making citizens part of the action" (Ruston and Kogut 2021), which has developed an action plan for collaborative decision-making based on replicas of virtual cities and huge volumes of online data and metadata generated by citizens.

4.2 Augmented Democracy: An Approach Based on Digital Twins

Augmented democracy refers to the use of smart technology —blockchain, digital twins, virtual politicians, non-fungible tokens, the metaverse, and so on— to improve the range of democratic processes in a broad sense, in terms of participation, security, robustness, control, traceability, electoral fraud, political corruption, and so on (Bernholz, Landemore and Reich 2021; Burgess 2021; Calvo 2020; Gordon 2020; Hidalgo 2018; Hidalgo and Calvo 2021b; MacKenzie 2021; Möller and Lange 2021; Navarrete et al. 2023; Pournaras 2020).

Paul Burgess (2021) has suggested four ways in which *blockchain*

technology and artificially intelligent mathematical models can be used to enhance or improve democratic processes: (1) the simple online transfer of a vote for representatives through the use of *blockchain* technology; (2) the use of online voting to approve or reject proposed changes to the legislature through *blockchain* technology; (3) the use of (anonymous) individual preferences to directly influence legislative decision-making through Artificial intelligence algorithms; and (4) the complete replacement of the (natural) legislature and the real people represented in it by a legislature composed of Artificial intelligence algorithms representing the voting public. Regarding the most radical form of augmented democracy, the replacement of legislators by AI algorithms, Burgess (2021: 99) adds that this is linked to the increasingly widespread idea that a "technocracy" operated by intelligent machines and based on huge volumes of data and metadata concerning the preferences, opinions, and actions of the electorate can outperform today's human-operated representative democracy, as it accounts for the needs of the population while avoiding the opaque motives of legislators.

> The most extreme idea of augmented democracy—which suggests the replacement of human legislators with algorithms—results in greater democracy and rule of law benefits than the other less extreme ideas. However, this article is not intended to be an argument for the replacement of human legislators with algorithms. My point is more tentative: even the most extreme innovations can enhance basic principles that we may fundamentally value; accordingly, even the most extreme ideas, notwithstanding intuitive responses, merit both continued and serious consideration (Burgess 2021: 99).

Nevertheless, Burgess's augmented democracy is an empty proposition, as it does not go beyond a mere simplified thought experiment on the positive impact that could be obtained by applying and using *blockchain* and smart algorithmic technology in modern democracies. The problem with Burgess's proposal is that it neither offers an in-depth exploration of the two technologies he considers appropriate for the emergence and development of an augmented democracy, nor does it show any concern for taking advantage of ideas set out in earlier, much more specific proposals for augmented democracy.

In this regard, one of the most interesting and influential proposals for augmented democracy is the one presented by César A. Hidalgo in 2018 (Hidalgo and Calvo 2021a; Calvo 2020, 2022). Hidalgo has proposed the application and use of digital twin technology for the creation and development

of this type of democracy. For Hidalgo, the use of virtual twins in politics would provide the possibility of minimizing and even eradicating the perennial and emerging problems that undermine the potential development of representative democracy —including political corruption, electoral fraud, highly ideological decision-making processes and the major lack of citizen participation— as well as the societies that employ this system as a means of organization. Hidalgo has argued that today "We have the tools to empower citizens and promote their participation. We could avoid many cases of corruption or poorly designed or unnecessary projects" (quoted in Sáez, 28 June 2018).

Likewise, Digital Twin technology would allow the enhancement of a political system through more objective decision-making and law, strategy and policymaking, which, in turn, would have a broader impact, as they draw on mass volumes of data and metadata concerning the preferences, opinions and decisions of society's hyperconnected corpus of individuals.

> The idea of augmented democracy is never to offload the democratic process to a senate of digital twins but to empower people's ability to participate by technologically expanding their cognitive bandwidth. By doing so, we may generate more instances of informed debate, not less. Hopefully, having a better user interface for democracy, and providing people with the ability to participate, will encourage people to become more aware of the regulations being discussed (Collective Learning Group 2022).

To achieve this, Hidalgo proposes moving towards an "automated direct democracy", which would be structured around a specialized user interface: a cyber-physical ecosystem, which permits the hyperconnectivity of citizens and the datafication of their preferences and opinions in conjunction with digital twins of the electorate, which in turn compresses and hyperlinks politics and citizenship. This system will be able to absorb a broad range of the electorate's voices in order to draw up policies and make political decisions linked to what citizens are, think and want (Hidalgo 2018; Sáez 26 June 2018, Collective Learning Group 2022; Hidalgo and Calvo 2021a; Calvo 2020, 2022). As Hidalgo explains,

> Imagine, for a second, a world in which, instead of having a representative that represents you and millions of other people, you can have a representative that represents only you, with your nuanced political views—that weird combination of libertarian and liberal and

> maybe a little bit conservative on some issues and maybe very progressive on others. Politicians nowadays are packages, and they're full of compromises. But you might have someone that can represent only you, if you are willing to give up the idea that representative is a human. If that representative is a software agent, we could have a senate that has as many senators as we have citizens. And those senators are going to be able to read every bill and they're going to be able to vote on each one of them. (Hidalgo 2018, 7: 55–8: 47).

Hidalgo stresses the need to take into account three important issues concerning the practical application and development of his proposal for an augmented democracy. Firstly, he highlights the need to both overcome the cognitive bandwidth problem of today's politicians while also eradicating corruption, nepotism, the embezzlement of public funds, influence peddling and other cases of political malpractice that prevent or hinder the development of democratic systems and, therefore, of modern societies. To do so, he believes it is necessary to broaden the information base underpinning public policymaking and political decision-making through Artificial intelligence algorithms. Endowed with a cognitive bandwidth capable of corresponding to the current hyperdigital technological moment, these algorithms would be able to operate with guarantees for the collection, processing and conversion of the massive volumes of data and metadata concerning the opinions, habits, behaviours and preferences of the digitally hyperconnected citizens, firstly, into relevant information and then, into knowledge that can be applied in political processes.

Secondly, Hidalgo proposes correcting the weak and indirect form of citizen participation in modern politics —one of the issues that is most often criticized by the social movements emerging within representative democracies and who demand it be redressed— through the design, application and use of personalized digital twins that collect, analyse and process data on the opinions, habits, behaviours and personal preferences of users of the democratic system.

Finally, Hidalgo points to the need to close the gap between what politicians say and do, and what society rightly expects from them by keeping governments hyperconnected with citizens via algorithm-based Artificial intelligence that can analyse the data and metadata used by public authorities, and the digital twins that recreate the users, institutions and organizations of the democratic system.

To obtain satisfactory results for the application of augmented democracy on each of these three fronts it has potential design principles and implementation requirements underpinned by a range of sources and studies. These consist of different approaches to orientating and formalizing actions for the design,

application and practical implementation of augmented democracy, but all are interconnected, convergent, complementary and scalable (Hidalgo 2015, 2018; Hidalgo and Calvo 2021a; Hidalgo et al. 2020; Sáez 2018; Collective Learning Group 2022; Alemany 2018; Luna and Pérez 2018; Calvo 2020; McInnes 2019; Navarrete et al. 2023; Pournaras 2020; Campos et al. 2019).

4.3 Design Principles and Implementation Requirements for Augmented Democracy

Augmented democracy is, first and foremost, an idea for the design, implementation and use of digital twins to expand citizens' capacity to participate in a high volume of democratic decisions, and to establish a mechanism capable of testing laws, decrees, policies and government strategies before they are implemented. In other words, it seeks to use virtual simulations to stress-test laws, decrees, policies or strategies by subjecting them to potential political, cultural, economic, social or natural factors or scenarios, such as a terrorist attack, a pandemic, the reaction from a social movement or a natural catastrophe. In this way it can test their resilience as well as potential, weaknesses, strengths and threats.

A digital twin, as a personalized virtual representation of a citizen or a democratic institution or organization, such as a political party, can be used to provide relevant information to its physical counterpart on decision-making issues or, directly, participate in decision-making processes on behalf of the natural person it represents. Therefore, as Hidalgo proposes, it would be a matter of generating "(…) an algorithm that learns from you, from your preferences, that you train, that is supplied with your data and predicts your choice (…)" (Hidalgo, quoted in Sáez 2018). And, just as the ideas of the natural person it represents evolves, the algorithm evolves too.

In the case of *augmented democracy*, users of an augmented democratic system log on to a digital platform, register as users and select one of many tools equipped with Artificial intelligence in order to create a digital representation (avatar) of themselves, which would be based on the behavioural data and metadata they produce. Through such digital twins, which would make personalized predictions or suggestions about their users' political decisions, the natural person they represent would be digitally hyperconnected to a digital forum, congress or parliament at either a local, regional or national level. This is where decrees, regulations, laws, projects, policies, and so on, would be analysed, debated and voted upon. The digital twin would participate within the

latter spaces as a personal and non-transferable political representative of the natural person, and their data would be combined with information related to the issue being analysed in order to predict what the natural person would vote or think about the issue, and thereby act accordingly. As Hidalgo states, "(…) we could have a Senate that has as many senators as citizens" (Hidalgo 2018).

In this sense, the complementary nature and synergic convergence of digital twins and advanced generative AI, especially stochastic algorithms and LLMs makes it possible to give augmented democracy a greater degree of specificity, precision, development and possibility, as well as glimpsing a definite route towards practical application.

As has been mentioned, generative AI uses big databases to process human content and extract patterns, relationships and statistical probabilities it then applies to generate new content, such as text, videos, audios and ultra-realist images. The level of development achieved in the last few years, especially concerning LLMs, allows generative AI to maintain interaction with humans via various supports and devices. This could be through implementation in robots or inclusion in digital platforms, websites, chats, social media and so on. Using these, generative AI maintains conversations with humans, analyses and summarizes texts, creates audiovisual content, draws up reports, and answers questions and requests of all kinds, among other things.

Currently, the synergic convergence of these two technologies–digital twins and generative AI–is already being developed and applied in different fields, such as administration, productive industry, the environment and health. And, as can be deduced from the different studies in this respect (Emmert 2023; Bordukova et al. 2023; Nyholm 2023; Booyse et al. 2020), this convergence offers great mutual possibilities for improving their development, functionality, potential and practical applicability.

In an *augmented democracy*, the implementation and use of *LLMs* in digital twins would allow closer, more fluid political communication that is more human and can be developed and understood by all parties involved and affected, with a positive influence both on the interaction between digital replicas– *Generative Digital Twins*–within the virtual forum or senate, and also between these and the physical entities they represent.

However, the possibilities of an *augmented democracy* based on digital twins are not limited to Hidalgo's proposed improvement of political participation processes. Quite the reverse, digital twin technology also makes it possible to use computational methods to replicate both the architecture and overall operation of a country's democracy, or a particular part of it, such as the electoral system, one of the three branches of state power—the legislature, executive, and

judiciary—or any one government institution, in order to test it under a variety of stress scenarios. The digital twin of a democracy therefore offers the possibility of anticipating the impact, for example, of the implementation of a new law or policy. This would also allow the gathering of information concerning the political system's robustness and capacity to respond to natural disasters, social change and ideological movements so that the model could be adjusted to avoid, as far as possible, any unintended consequences that might harm society or the democratic system itself (García-Marzá and Calvo 2022).

To create and refine this macro-democratic digital twin of part or a whole democratic system, augmented democracy is supported and enhanced by the system's micro-democratic and meso-democratic digital twins. These are the digital representations generated, firstly, by the different institutions and organizations making up a country's democratic structure (meso-democratic level) and, secondly, by the range of people making up that country's democratic citizenry (micro-democratic level).

Also, as happens with automated political participation, the digital twins of part or all of a democratic system can be provided with a higher level of potential, predictability, intelligibility and accessibility through their synergic convergence with generative AI, particularly via LLMs and the Generative Adversarial Network (GAN) (Zotov et al. 2020; Quilodrán-Casas et al. 2022; Li and Yang 2021).

GANs produce groups of data or content using two antagonistic algorithms: one that produces while the other discriminates. A convolutional neural network produces content with realistic synthetic data based on the characteristics and patters of, for example, a large set of images. Meanwhile, a transposed convolutional neural network analyses the content produced by the generative algorithm to distinguish which data are real and which are false (artificially generated). Their common aim is to generate groups of data where it is almost impossible to distinguish or discriminate the real from the synthetic. In this way, the antagonistic way these algorithms are complemented through this zero-sum game allows the constant, innovative development of the datasets produced by the GAN.

In the case of augmented democracy, the inclusion and use of GANs and LLMs would make it possible to develop, improve and innovate continuously and in real time both in terms of the architecture and the design and general operation of democracy or a specific part of it. Firstly, this would be by offering new ideas and solutions to improve democratic systems, designs and processes. Secondly, it would generate self-improvement and self-innovation processes for the virtual replica of the systems and digital processes. Thirdly, it would provide

predictions on possible anomalies and risks of the systems and processes. Finally, via the possibility that the people responsible for the democratic systems could interact and maintain direct two-way communication with the digital replicas of the designs to discover, request and resolve queries about different elements of their structure and operation.

Both possible applications of augmented democracy, digitally replicate citizens to produce automated direct participation in democratic processes and digitally replicate a democratic system or parts of it to virtually experiment with innovative ideas or possible cultural or natural changes, have already been tested on a municipal scale in the project "Digital Twins: Making citizens part of the action" (Ruston and Kogut 2021), funded by the European Commission. Firstly, at a municipal level, the project seeks to offer a pilot scheme in the form of a digital twin for smart cities. This will be based on the data and metadata generated within urban space, the deployment of the digital twin on a 3D interface and the principle of explicability (this AI concept is discussed in depth in Chap. 10) to endow the system with the reliability it needs. It is intended to bring about a substantial improvement in the level and quality of citizen participation in municipal policymaking and political decision-making processes. Secondly, the project aims to provide virtual replicas of cities to provide a clearer understanding of the complex interrelationship between traffic, air quality, noise and other urban factors, as well as a means of analysing the expected impact of potential changes in order to make better evidence-based operational decisions. However, this project does not contemplate the possibility of including and using generative AI, such as GANs and LLMs to improve the design, development, robustness and applicability of the model.

Yet, aside from the medium which allows augmented democracy to be implemented –digital twin technology– along with its possibilities –citizen participation and analysis, the detection and prevention of the impact produced or likely to be produced by changes in a broad sense– the practical application of an augmented democratic system must be based and developed with regard to at least nine design principles and applicability requirements: *dataficability*, hyperconnectivity, scalability, predictability, representativeness, scrutability, accessibility, collectivity and synchronicity, as well as any application requirements. These can be summarized as follows:

– *Dataficability*: augmented democratic systems rely on massive volumes of online data and metadata concerning the preferences, habits, opinions, trends and behaviour generated by the real individuals they represent, as well as legal entities, such as institutions and organizations. Diverse digital data

conversion and analysis techniques and technologies will be used, such as data mining techniques including association rule learning and cluster analysis.

– *Hyperconnectivity:* augmented democratic systems are based on the digital self-hyperconnectivity of the natural persons represented by the digital twins, which makes it possible to establish a kind of permanent connection between the physical world and the virtual world, and thereby generate a form of collector and processor of personal or collective information. To do so, the natural person must choose and use one or more of the available hyperconnectivity tools, such as questionnaires (on reading habits, political opinions, consumer preferences, etc.), mobile devices, or radio frequency identification chips grafted under the skin (RFID), which are capable of collecting and digitizing continuously and in real time the massive volumes of data and metadata generated by the person concerning their opinions, habits, behaviours, trends and preferences.
– *Scalability*: Augmented democratic systems rely on scalable and distributed information generation and deliberation processes distributed by mapping out high-dimensional opinion spaces; these enable it to overcome the limitations of small group meetings. For this purpose, scaled physical and digital representations are introduced into participation and deliberation platforms, such as *Polis* (Small et al. 2021). They are capable of collecting, analysing and synthesizing people's feedback in a scalable and distributed way in order to generate a shared understanding, discourage counterproductive behaviour (trolling) and cultivate points of consensus (Small et al. 2021).
– *Predictability:* augmented democratic systems rely on the diagnostic and prognostic capabilities of digital twins. This capacity permits the anticipation of the needs and demands of the natural persons represented for decision-making and the drafting of regulations, laws, proposals, projects, strategies, policies, and so on. Furthermore, it can forecast their short, medium and long-term impact. To achieve this, digital twins are currently complemented with generative AI, such as LLMs and GANs.
– Representativeness: augmented democratic systems are based on the virtual representativeness of real citizens. Therefore, the natural persons represented by the digital twins must be able to choose which one of the various technological and technical possibilities best represents them, and is most appropriate for their interests and expectations.
– *Scrutability:* augmented democratic systems are based on reinforcement learning programmes, under which the individuals they represent would be able to access the prognoses and diagnoses made, as well as the decisions and

suggestions made by their digital twins within the spaces of political participation, for the purpose of assessing, correcting, refining and/or reinforcing the digital twins' activity. Augmented democratic systems will therefore need to design and offer simple and accessible reinforcement learning tools, as well as implementing training programmes for their use which are appropriate for the needs and skills of the entire democratic community.

– *Accessibility:* augmented democratic systems are based on accessibility to technology and the techniques they are based on and developed with. Thus, the tools needed by the individual represented by the virtual twin must be accessible and appropriate to their level of technological knowledge.
– *Collectivity:* enhanced democratic systems are enriched and developed through collective learning, i.e. through collective processes of knowledge creation and dissemination among countries, institutions, organizations and individuals based on experience, similitude or relatedness and knowledge intensity. Therefore, strategies and channels for knowledge dissemination must be developed along with tools for large-scale data integration, distribution, flow and visualization (Hidalgo 2015).
– *Synchronicity*: augmented democratic systems are based on the hybridization between the virtual and real worlds. To do this, it is necessary to provide the system with hybridization technologies allowing links in real time between the real and digital democratic spaces.

The application of digital twins along with these nine design principles or applicability requirements would enable the emergence and implementation of an augmented democracy, with the main aim of increasing and improving civic participation and influence on politics, as well as anticipating the impact and consequences of both changing and implementing public laws, strategies and policies.

Nevertheless, despite the convincing dimensions of this proposal, many issues remain to be fully developed before any actual implementation of an augmented democracy. For example, it remains to be decided who will draft the decrees, laws, projects, policies, proposals, strategies, and so on to be analysed, discussed and voted on by the digital twins in the virtual participation spaces? How long will it take to guarantee universal access to the system, the security of users' data and equal opportunities? And finally, how, among many other things, can interference by the powers that be in the system be avoided?

My goal is to ignite a debate focused on expanding the design space of

> democracy. Democracy is a beautiful idea, but there are many ways in which it can be implemented. The number of possible implementations of democracy has increased with the introduction of new technologies, but we have not been diligent at exploring many of these alternatives. My proposal is to create unofficial experimental systems of augmented democracy that we can use to learn about the limits and potential of this technology (Collective Learning Group 2022).

In short, augmented democracy has become tremendously important as a means of addressing the traditional critiques of modern democracies while at the same time eradicating the failings that hinder their development and have a negative impact on societies that depend on these political systems. However, for this to happen, as Hidalgo explicitly and implicitly recognizes, work must be undertaken on the strategic-technological dimension of the design and application of augmented democracy, as well as its emotional and moral dimensions to ensure so that it is not only economically and technically viable, but also possible, valid and desirable.

> I think AI ethics is an important topic, primarily because we know little about it. In recent years many opinion pieces, editorials, and essays have been written about AI ethics. But there is still relatively little data on the topic, except for a few noteworthy examples in the case of self-driving vehicles or computer vision bias (Collective Learning Group 2022).

4.4 The Ethical and Political Challenges of *Augmented Democracy*

As Pedro Baños argues in *La encrucijada mundial. Un manual del mañana* [*The global crossroads. A handbook for tomorrow*] (2022), in this new datafied, hyperdigitized, algorithmized and robotized world we are creating "The heart of the issue is, as it is for everything, achieving the balance between technology and democracy. To do this, we would have to work out how to get the most from the technology through automation processes or by using Big Data to be closer to citizens and really meet their true needs and interests but without sacrificing democratic values and principles" (2023: 129). However, there are more than a few ethical and political problems arising as a result of digital transformation and they are not easy to solve.

Among the range of ethical and political issues detected in the current use of

Artificial intelligence algorithms in the political sphere (Calvo 2019), there has been a considerable increase in:

(a) social exclusion linked to the degree of citizens' hyperconnectivity;

(b) heteronomy and civic paternalism;

(c) biased decision-making of a homophobic, aporophobic, xenophobic and/or misogynist nature;

(d) intrusion into the private sphere of groups and individuals;

(e) the opacity of intelligent systems and processes;

(f) the distortion and manipulation of public opinion through social bots;

(g) the effectiveness of tools for intercepting and obstructing political communication;

(h) the emergence of psychosocial pathologies linked to surveillance and an excessive use of control;

(i) and a decline in efforts to implement policies ensuring accountability and responsibility for actions and decisions.

In this regard, the case of the *Bosco* algorithm used by the Spanish government to grant financial support to prevent fuel poverty is highly illustrative (Ollero 2019). As reported in 2019 by the *Fundación Civio*, this mathematical model denied aid applications from people who met all the requirements stipulated during the application process. Even more seriously, and despite the evidence and the consequences it had for the most vulnerable members of the democratic system, it has not been possible to confirm the causes behind these anomalies in order to correct them. In 2019 the Spanish state invoked an alleged intellectual property issue to deny experts access to the algorithm's source code, and, more recently, in 2022 it has refused access on the grounds that it would breach public security (Pascual 2022).

There are also significant cases of the use and abuse of mass surveillance algorithms by state governments. On one hand, the *Social Credit* programme launched by the Chinese government merits careful scrutiny. The installation of

over 600 million cameras and sensors—hyperlinked to facial recognition software (digital phrenology) and *Big Data* analysis—in streets, buildings, public transport, schools and inside rented homes has been set up and funded for the purpose of detecting and tagging uncivil, unseemly or criminal behaviour among citizens (O'Neal 2016; Espinosa 2017; Frank 2019). On the other hand, the multiple errors detected in the use of Artificial intelligence for detecting criminals and preventing crime through the use of facial recognition and other data analysis techniques are likewise noteworthy. For example, the system used by the Chicago police to rank the 400 people most likely to commit violent crimes in the future included citizens with no prior criminal record (O'Neal 2016); the facial recognition system implemented in the city of Buenos Aires to trace criminals had an extremely high error rate (Bercovich 2019); and, finally, the *VioGen* algorithm used by the Spanish police to assess the risk faced by domestic violence victims, despite having been thoroughly revised on several occasions since 2014 following the detection of serious errors in the assessments and forecasts it made (Álvarez and Sanmartín 2014), was once again involved in a major scandal in 2022 (Amantegui 2022).

Finally, as will be seen below, the fraudulent use of *political bots* by governments and political parties to manipulate public opinion and interfere in political communication for the purpose of soliciting contributions or votes, feigning greater popularity on social media, disrupting the communication strategy of rivals or launching false opinions to influence the electorate, has become notorious. For example, the former US President Barack Obama flooded social media with automated messages in an attempt to attract public attention and support; Mitt Romney was accused of buying thousands of *Twitter* followers in an attempt to appear more popular; and Donald Trump spent 70 million dollars on Facebook advertisements to improve his image and attack his opponents; he also paid for social bots to launch favourable opinions in forums in addition to fake profiles intended to increase his followers (Howard et al. 2018).

For these and other reasons, the design, emergence and success of *augmented democracy* depends not only on its scientific and technical development, but also on its moral and emotional development. Hidalgo (2015, 2018, 2020; Sáez 2018, Calvo 2020; Hidalgo and Calvo 2021a; Collective Learning Group 2022) is wholly aware of the current scientific-technological, moral and emotional immaturity of any algorithmic proposal for democracy and the negative impact that its hasty application would have on society. It is therefore necessary to work on the implementation requirements for *augmented democracy* by, for example, recreating a *cybernetic ecosystem* that would be capable of ensuring: the moral

validity of the design and practical implementation of an augmented democracy proposal; appropriate behaviour by companies, organizations, users and smart devices; and, finally, the economic, social and environmental impact of the augmented democratic system.

A *cyberethical ecosystem* would comprise all the elements, processes, mechanisms and factors involved in recreating and implementing an environment for the communication and deliberation needed to manage, monitor and enforce ethics in the practical sphere, and be capable of responding to the current challenges raised by digitization in diverse fields and activities (Calvo 2020).

In terms of *augmented democracy*, among other things a *cyberethical ecosystem* would, on the one hand, be concerned with specifying the different mechanisms and instruments available to those involved and affected by the system for communication with and between them; public scrutiny of its design, the behaviour it implies and its eventual impact; the management of notifications concerning suggested improvements to the system and its processes; discussing and resolving conflicts underpinning immoral behaviour and non-compliance with norms and values and commitments reached; detecting and eradicating any negative consequences arising from the practical implementation of augmented democracy.

However, on the other hand, a *cyberethical ecosystem* would address the unravelling of the physical, technological, emotional and moral factors involved in the emergence, development, supervision and survival of the democratic system from a just and responsible point of view. These would include the degree of entropy it supports, the volume, complexity and flow of information available and its possible increase over the short, medium and long-term; the system's technical, natural, cultural, emotional and deontological limitations; the moral horizon that guides actions and decisions; the emotional elements underlying citizens' participation in the proposed political process (at the very least, the datafication of opinions, behaviours, habits and preferences, inclusion in the system and the training of the *digital twin*). It would also cover the goals pursued and the virtues involved in obtaining them. Finally, it would specify the people, groups or AI devices that could affect or be affected by the system, their interests and their power and responsibility for the impacts generated.

Concerning the moral dimension of the development of augmented democracy, a *cyberethical ecosystem* should recreate an ethical infrastructure consisting of different mechanisms for managing and enforcing its ethics, and this would involve: a Code of Ethics and Conduct, an ethics committee, ethics audits, ethics lines**,** accountability reports and explicability updates (García-

Marzá 2017; García-Marzá and Calvo 2022; Calvo 2020, 2021, 2022). The principal reason for introducing these is to enable the emergence and empowerment of an inclusive, egalitarian, symmetrical and open space for dialogue, which fosters and tends towards intersubjective agreement on a range of issues among all persons or groups who may affect or be affected by an augmented political process. This ethical infrastructure is explored in the final chapters of this book.

Secondly, with regard to the moral dimension of the development of *augmented democracy*, a *cyberethical ecosystem* should be supported by *blockchain* technology. As will be seen in Chap. 12, this would enable the traceability, comparability, scrutability and auditability of the information flow underpinning the system, which would support its potential and sustainability. In particular this concerns the system's design; the source code of the different algorithms used; the opinions, habits, behaviours and political preferences for each of the users of the system, which are collected, analysed and processed by the *digital twins*; the policies, laws, decrees, proposals, amendments and projects elaborated and deliberated in the different democratic institutions; public contracts and political decisions; and, finally, public aid. It would also engage with reports on the economic, social and environmental impacts of government activity; codes of ethics and conduct, good practice, good governance and/or professional ethics; social actions and policies; international guidelines and so on.

Furthermore, and just as importantly, a *cyberethical ecosystem* supported by *blockchain* technology would be able to guarantee users' confidentiality, especially *whistleblowers* who can interact in the system via the ethics hotline to enhance or safeguard it with suggestions and proposals for improvement of all matters related to ethical issues or alerts or complaints of professional malpractice.

Finally, in order to develop the moral dimension of augmented *democracy* a *cyberethical ecosystem* must have: a system for monitoring the social and moral impact of actions and decisions based on *Big Data* technology—tools capable of analysing the huge volumes of data selected for specific or general aspects of the ethical and socially responsible dimension of a democracy based on algorithms; *cyber-blowers* or *cyberethical alerts;* Artificial intelligence algorithms that can detect anomalies *(outliers*) capable of negatively affecting the system's integrity and social development and the committed participation of civil society; and, finally, the generation of data on opinions, assessments, preferences and/or *criticisms* of the system obtained through digital hyperconnectivity devices, which enable the collection, recording and processing of data to ascertain the

degree of consensus on different issues.

To conclude, the multiple problems that have plagued modern democratic systems since their emergence in the late eighteenth century, such as corruption, nepotism, influence peddling, embezzlement of public funds, a lack of professionalism and public commitment, opacity, and so on, together with the emergence of the Digital Age and its consequences—the digital hyperconnectivity of society and the possibility of the datafication of the social, subjective as well as objective worlds—have generated multiple disruptive proposals for democratic models based on algorithms that make objective decisions based on the common good or the opinion of the majority of citizens.

Beyond the radical, poorly developed and by no means consensual proposals sponsored by technology companies and entrepreneurs, which seek to insert Artificial intelligence algorithms into local, regional and national government in countries such as Japan, New Zealand, Russia, Romania and Denmark. Other proposals are undertaking both a theoretical and practical collective development over the medium or long-term. Hidalgo's augmented democracy proposal is a key example of this. Far from presenting a finished design, it seeks to provide a starting point for the establishment of an interdisciplinary and critical dialogue on design, application and use of his system.

The proposed integration of a *cyberethical ecosystem* into Hidalgo's model of augmented democracy would undertake the management, monitoring and fulfilment of its ethical dimensions and ensure that a regulatory framework guides the actions taken and decisions made by the political representatives, as well as the rest of the system's users. It would also uphold the excellent character of the professionals involved in its operation, offer public scrutiny of the processes and provide a basis for the intersubjective agreement of all those affected by the current or future consequences of the practical application of *augmented democracy*, while also ensuring that all the commitments made by the system are fulfilled, and that there is critical reflection on the actions and decisions derived from them.

Bibliography

Alemany, A. 2018. Un físico del MIT propone cambiar a los políticos por un sistema de Inteligencia Artificial. *Perspectiva*. https://diarioperspectiva.com/un-fisico-del-mit-propone-cambiar-a-los-politicos-por-un-sistema-de-inteligencia-artificial/.

Álvarez, R.J., and O.R. Sanmartin. 2014. El test que valora el riesgo de las maltratadas será modificado. *El Mundo*. Available at: https://www.elmundo.es/espana/2014/12/10/54875ceaca47412f558b4586.html.

Amantegui, A. 2022. El algoritmo que evalúa el riesgo de las víctimas de violencia de género no funciona como debería. *La Vanguardia.* Available at https://www.lavanguardia.com/tecnologia/20220314/8123126/algoritmo-evalua-riesgo-victimas-violencia-genero-funciona-deberia-pmv.html.

Asimov, Isaac. 1955. Franchise. *If. Worlds of Science Fiction* (nº5, vol. 5). New York: Quinn Publishing.

Baños, Pedro. 2022. *La encrucijada mundial. Un manual del mañana*. Madrid: Ariel.

Baños, Pedro. 2023. *La encrucijada mundial. Un manual del mañana*. Madrid: Ariel.

Bercovich, F. 2019. Vigilar y equivocarse. *Cenital.* Available at: https://www.cenital.com/2019/08/13/vigilar-y-equivocarse/64022.

Bernholz, L., H. Landemore, and R. Reich. 2021. *Digital Technology and Democratic Theory*. Chicago: University of Chicago Press.
[Crossref]

Booyse, Wihan, Daniel N. Wilke, and Stephan Heyns. 2020. Deep digital twins for detection, diagnostics and prognostics. *Mechanical Systems and Signal Processing* 140: 106612.
[Crossref]

Bordukova, Maria, Makarov, Nikita, Rodriguez-Esteban, Raul; Schmich, Fabian, and Menden, Michael P. (2023). Generative Artificial intelligence empowers digital twins in drug discovery and clinical trials. *Expert Opinion on Drug Discovery*, First on-line: 1–10. https://doi.org/10.1080/17460441.2023.2273839.

Burgess, P. 2021. Algorithmic Augmentation of Democracy: Considering Whether Technology Can Enhance the Concepts of Democracy and the Rule of Law Through Four Hypotheticals. *AI & Society* 37: 97–112. https://doi.org/10.1007/s00146-021-01170-8.
[Crossref]

Calvo, P. 2019. Democracia algorítmica: Consideraciones éticas sobre la *dataficación* de la esfera pública. *CLAD Journal. Reforma y Democracia* 74: 5–30.

———. 2020. Democracia aumentada: un *ecosistema ciberético* para la participación política basada en algoritmos. *Apeiron. Revista de Filosofía* (12): 129–141.

———. 2021. El gobierno ético de los datos masivos. *Dilemata. Revista de Éticas Aplicadas* 34: 31–49.

———. 2022. Gemelos digitales y Democracia. *CLAD. Reforma y Democracia. Revista de Éticas Aplicadas* (83).

Campos, P., A. Ramírez, G. Saavedra, M. Uribe, and L. Valladares. 2019. iVote. Available at https://static1.squarespace.com/static/5c2e56155ffd20ae8ee2c3e4/t/5d77c7e848eb396542e3a6ca/1568131061767/iVoteAD.pdf.

Collective Learning Group. 2021. *Augmented Democracy. Exploring the Design Space of Collective Decisions*. Toulouse, Center for Collective Learning, https://www.peopledemocracy.com.

———. 2022. *Augmented Democracy. Exploring the design space of Collective decisions*. Available at: https://www.peopledemocracy.com/.

Cox, R. 2020. Factors Affecting Conversation Design for Digital Political Personae. *Doctoral thesis*.

Available at https://researcharchive.vuw.ac.nz/xmlui/bitstream/handle/10063/9168/thesis_access.pdf?sequence=1.

Dassault, S. 2022. *The Living Heart Project. A Translational Research Initiative to Revolutionize Cardiovascular Science Through Realistic Simulation*. Available at https://www.3ds.com/products-services/simulia/solutions/life-sciences-healthcare/the-living-heart-project/.

Emmert, Frank. 2023. What Is the Role of AI for Digital Twins? *AI 4* (3): 721–728.
[Crossref]

Ericsson, H., and A.D. Little. 2020. *Connected Manufacturing: A Guide to Industry 4.0 Transformation with Private Cellular Technology*. Available at https://www.hannovermesse.de/apollo/hannover_messe_2021/obs/Binary/A1089023/Connected%20Manufacturing%20Report_compressed.pdf.

Espinosa, J. 2017. Algoritmos y 600 millones de cámaras: así funcionará el 'carné de buen ciudadano' chino. *El Mundo*. Available at: https://www.elmundo.es/papel/historias/2017/12/15/5a327dece5fdea34758b45ef.html.

European Commission. 2020. *Destination Earth* [European strategy for data] Brussels. https://digital-strategy.ec.europa.eu/en/policies/destination-earth.

Frank. 2019. Chinese Authorities Require Surveillance Cameras to Be Installed Inside Rental Housing. *The Epoch Times*. https://www.theepochtimes.com/china/chinese-authorities-require-surveillance-cameras-installed-inside-rental-housing-2875997.

García-Marzá, D. 2017. From Ethical Codes to Ethical Auditing: An Ethical Infrastructure for Social Responsibility Communication. *El Profesional de la Información* 26 (2): 268–276.
[Crossref]

García-Marzá, D., and P. Calvo. 2022. *Algorithmic Democracy: A Critical Perspective from Deliberative Democracy*. Cham: Springer.

Gerach, T., S. Schuler, J. Fröhlich, L. Lindner, E. Kovacheva, R. Moss, E.M. Wülfers, G. Seemann, Ch. Wieners, and A. Loewe. 2021. Electro-Mechanical Whole-Heart Digital Twins: A Fully Coupled Multi-physics Approach. *Mathematics* 9 (11): 1247. https://doi.org/10.3390/math9111247.
[Crossref]

Glaessgen, E., and Stargel, D. 2012. The Digital Twin Paradigm for Future NASA and US Air Force Vehicles. Proceedings of the 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA; 2012 Apr 23–26; Honolulu, HI, USA, American Institute of Aeronautics and Astronautics, Reston, p. 1818.

Gordon, J.S., ed. 2020. *Smart Technologies and Fundamental Rights*. Brill: Leiden.

Grieves, M. 2014. Digital Twin: Manufacturing Excellence Through Virtual Factory Replication. White paper US Florida Institute of Technology, Melbourne.

Hidalgo, C.A. 2015. *Why Information Grows: The Evolution of Order, from Atoms to Economies*. New York: Basic books.

———. 2018. *A Bold Idea to Replace Politicians*. Vancouver: TED. https://www.ted.com/talks/cesar_hidalgo_a_bold_idea_to_replace_politicians#t-11731.

Hidalgo, C.A., and Patrici Calvo. 2021a. The Dawn of Digital Democracy. In *Democracy in the Age of Artificial Intelligence*, 1–3. Castellón: Universitat Jaume I.

———. 2021b. The Dawn of Digital Democracy. *Recerca. Revista de Pensament i Anàlisi* 26 (2): 1–3. https://doi.org/10.6035/recerca.6113.
[Crossref]

Hidalgo, C.A., D. Orguian, J. Albo-Canals, F. de Almeida, and N. Martín. 2020. *How Humans Judge Machines*. Cambridge: The MIT Press.

Himmelreich, J. 2020. Should We Automate Democracy? In *Oxford Handbook of Digital Ethics*. Oxford: Oxford University Press.

Howard, Ph.N., S. Woolley, and R. Calo. 2018. Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration. *Journal of Information Technology & Politics* 15 (2).

India.com. 2017. New Zealand Develops World's First Artificial Intelligence Politician. India.com. https://www.india.com/news/world/new-zealand-develops-worlds-first-artificial-intelligence-politician-2678383/.

Kamel Boulos, M.N., and P. Zhang. 2021. Digital Twins: From Personalized Medicine to Precision Public Health. *Journal of Personalized Medicine* 11 (8): 745. https://doi.org/10.3390/jpm11080745.
[Crossref]

Korenhof, P., V. Blok, and S. Kloppenburg. 2021. Steering Representations: Towards a Critical Understanding of Digital Twins. *Philosophy and Technology 34* (4): 1751–1773. https://doi.org/10.1007/s13347-021-00484-1.
[Crossref]

Langelaar, W. 2018. SAM (Semantic Analysis Machine). *6th Conference on Computation, Communication, Aesthetics & X*. Madrid, Spain. Available at: https://2018.xcoax.org/pdf/xCoAx2018-Langelaar.pdf.

———. 2020. Factors Affecting Conversation Design for Digital Political Personae. http://researcharchive.vuw.ac.nz/bitstream/handle/10063/9168/thesis_access.pdf?sequence=1.

Langelaar, W., N. Gerritsen, and A. Smith. 2017. Sam: Virtual Politician. *Scoop*. Available at http://business.scoop.co.nz/2017/11/15/meet-sam-the-virtual-politician-for-the-digital-age/.

Li, C., and D. Yang. 2021. Construction of Power Grid Digital Twin Model Based on GAN. *2021 China Automation Congress (CAC)*, Beijing, China, pp. 7767–7771. https://doi.org/10.1109/CAC53003.2021.9728190.

Luna, J.P., and C. Pérez 2018. ¿Democracia sin políticos? La engañosa fe en los algoritmos. *Ciper*. Available at https://ciperchile.cl/2018/05/09/democracia-sin-politicos-la-enganosa-fe-en-los-algoritmos/.

MacKenzie, M.K., ed. 2021. *Future Publics: Democracy, Deliberation, and Future-Regarding Collective Action*. Oxford University Press: Oxford.

McInnes, E. 2019. Theoretical Underpinnings of a Pure Digital Democracy Model of Government. Available at https://static1.squarespace.com/static/5c2e56155ffd20ae8ee2c3e4/t/5d77c81dedd5084e5c931575/1568131114349/Theoretical+Underpinnings+of+a+Pure+Digital+Democracy+-

+Preliminary+Version.pdf.

Möller, L., and D. Lange. 2021. *Augmented Democracy in der Politischen Bildung: Neue Herausforderungen der Digitalisierung*. Cham: Springer.
[Crossref]

Navarrete, C., M. Macedo, R. Colley, Zhang, N. Ferrada, M.E. Mello, R. Lira, C. Bastos-Filho, U. Grandi, J. Lang, and C.A. Hidalgo. 2023. Understanding Political Divisiveness Using Online Participation Data From the 2022 French and Brazilian Presidential Elections. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-023-01755-x.

Negri, E., L. Fumagalli, and M. Macchi. 2017. A Review of the Roles of Digital Twin in CPS-Based Production Systems. *Procedia Manufacturing* 11 (2017): 939–948.
[Crossref]

Nyholm, Sven. 2023. Is Academic Enhancement Possible by Means of Generative AI-Based Digital Twins? *The American Journal of Bioethics* 23 (10): 44–47.
[Crossref]

O'Connor, T. 2017. Will the Next Russian President be a Robot? Putin's New Challenger is a Machine that Knows 'Everything'. *Newsweek*. https://www.newsweek.com/russia-putin-could-face-controversial-robot-next-year-president-election-741509.

Ollero, D.J. 2019. The Government's Secret Algorithm that Decides Whether You Get a Subsidy for Your Electricity Bill. *The World*. https://www.elmundo.es/tecnologia/2019/07/03/5d1b89fbfc6c83a2358b46ca.html.

O'Neal, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

Papanagnou, C. 2019. A Digital Twin Model for Enhancing Performance Measurement in Assembly Lines. In *Digital Twin Technologies and Smart Cities*, ed. Maryam Farsi, Alireza Daneshkhah, Amin Hosseinian-Far, and Hamid Jahankhani, 53–66. Cham: Springer. https://doi.org/10.1007/978-3-030-18732-3_4.
[Crossref]

Pascual, M. 2022. Un algoritmo opaco al que han detectado errores decide quién recibe ayudas públicas del bono social. *El País*. Available at https://elpais.com/tecnologia/2022-04-22/un-algoritmo-opaco-al-que-han-detectado-errores-decide-quien"-recibe-ayudas-publicas-del-bono-social.html.

Pournaras, E. 2020. Proof of Witness Presence: Blockchain Consensus for Augmented Democracy in Smart Cities. *Journal of Parallel and Distributed Computing* 145: 160–175.
[Crossref]

Quilodrán-Casas, C., V.L.S. Silva, R. Arcucci, C.E. Heaney, Y. Guo, and C.C. Pain. 2022. Digital twins Based on Bidirectional LSTM and GAN for Modelling the COVID-19 Pandemic. *Neurocomputing* 22 (470): 11–28. https://doi.org/10.1016/j.neucom.2021.10.043.
[Crossref]

Ruston, S., and P. Kogut. 2021. *DUET: Digital Urban European Twins*. Brussels: European Commission.

Sáez, C. 2018. Augmented Democracy. *CCCB*. http://lab.cccb.org/en/democracia-augmentada/.

Sarimova, L. 2017. My Main Opponent is Ksenia Sobchak: Yandex's VOICE assistant to Run for President. *Realnoevremya*. https://realnoevremya.com/articles/2025-yandexs-voice-assistant-to-run-for-president.

Serikawa, Y. 2019. [AI policy is coming], Nikkei. Available at https://www.nikkei.com/article/DGXMZO41080150Y9A200C1TCR000/.

Small, C., M. Bjorkegren, T. Erkkilä, L. Shaw, and C. Megill. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *Research. Revista de Pensament I Anàlisi* 26 (2): 1–26. https://doi.org/10.6035/recerca.5516.
[Crossref]

Tao, F., J. Cheng, Q. Qi, M. Zhang, H. Zhang, and F. Sui. 2018. Digital Twin-Driven Product Design, Manufacturing and Service with Big Data. *The International Journal of Advanced Manufacturing Technology* 94 (9–12): 3563–3576.
[Crossref]

Tao, F., Q. Qi, L. Wang, and A.Y.C. Nee. 2019a. Digital Twins and Cyber-Physical Systems Toward Smart Manufacturing and Industry 4.0: Correlation and Comparison. *Engineering* 5 (4): 653–661.
[Crossref]

———. 2019b. *Digital Twin Driven Smart Manufacturing*. London: Academic Press.

Wagner, M. 2017. This Virtual Politician Wants to Run for Office. *CNN*. https://edition.cnn.com/2017/11/23/tech/first-virtual-politician-trnd/index.html.

Zaripov, R. 2017. Искусссственнный интелллект балллотируется в президенты [Artificial intelligence running for president]. https://youtu.be/on-1TPMyc50.

Zotov, E., A. Tiwari, and V. Kadirkamanathan. 2020. Towards a Digital Twin with Generative Adversarial Network Modelling of Machining Vibration. In: Iliadis, L., Angelov, P., Jayne, C., Pimenidis, E. (eds) *Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference*. EANN 2020. Proceedings of the International Neural Networks Society, vol 2. Springer, Cham. https://doi.org/10.1007/978-3-030-48791-1_14.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_5

5. Platformization: The Dangers of the Artificial Public Sphere

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

What interests us in our context is that the condition to exercise an act of common will can only be met in an inclusive public sphere… Public communication (öffentliche) forms the necessary nexus between the political autonomy of each individual and the formation of the common political will of all citizens.
Jürgen Habermas, Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik (2022, 89)

Abstract
The aim of this chapter is to contrast the concept of the public sphere, a basic pillar of any democratic system, with the reality and expectations of a digital society. Big data and metadata have become a double-edged sword for the digitally hyperconnected democratic society. Meanwhile, the incredible potential of Big Data and the different techniques and technologies for exploiting data and metadata make it a product coveted by a system of institutions making up both the state and civil society. Although algorithmic democracy is based not only on the public sphere, the aim is to show the incompatibility between democratic participation and a public sphere artificially fabricated by technology platforms and their algorithms. In this chapter, we will look at the importance of the public sphere for democracy and the technologies that are constructing what we might call an artificial public sphere. We will be guided by the concept of the public sphere drawn from Habermas's most recent works and discussions about whether we are looking at a new structural change in the public sphere or rather its fragmentation into a plurality of semi-public spheres now incapable of generating quality public opinion and, still less, digital citizenship.

Keywords Artificial public sphere – Big data – Cyber-physical ecosystems – Polítical social bots – Generative AI – Synthetic public opinión – Big tech

5.1 The Public Sphere as a Basic Requirement for Democracy

As with this type of abstract concept that is a true deposit of learning and summary more or less achieved from past experience, the concept of public opinion is not based on nothing. We can trace it back to the vox populi of the Roman Empire, the *consensus* of medieval doctrine, Machiavelli's public reputation, Kant's principle of publicity and Rousseau's general will, and so on. In all these we find traces of this basic function of public opinion as a source of democratic legitimacy, and a criterion for the legitimation of public power (Habermas 1990; Keane 2009).

Behind the concept of the public sphere is always the need for the rational acceptance of power; the need to be based on the will not of an individual or group, but rather on the rational, reasoned will of all people affected. Power, in any of its dimensions, demands justification, based on the expectation of credibility required for its use and the idea our concept transmits is that political power is a public power whose targets and effects are public and precise, as a result, public acceptance. Ultimately, it is the condition that supports the idea that the law must be a result of a rational agreement or consensus on the general interest. This is a *regulatory idea* requiring a space and its own conditions to be able to exercise a critical perspective. Effectively, we are looking at a normative concept resulting from the reconstruction of intuitive knowledge which, as citizens, we in fact use to define, maintain and develop the trust democratic decisions deserve from us. The mere counting of votes cannot guarantee these expectations. In the words of Habermas: "Even though they may be idealizations, which can exceed real practice to a greater or lesser degree, these expectations generate social facts in as far as they become real in the judgement and behaviour of citizens" (Habermas 2022, 17).

This connection between the public sphere and democracy is already clear in the development of the first bourgeois constitutions (Sartori 1996; Keane 2008), distinguishing clearly this principle of publicity as an organizational principle (García-Marzá 2012). It is none other than the basic function of parliamentary procedures to ensure the influx of public opinion, the relationship between representatives and voters, the first laws guaranteeing fundamental rights of freedom of assembly, expression and the press, and so on. The intrinsic and

necessary relationship between public opinion and democracy can be summarized in three steps:

– Democracy is based on popular sovereignty as a principle of legitimation; that which "we could all want".
– The people as sovereign have to know and say, they have to possess and be able to express and debate content, knowledge and reasons backing the possible justice of the established order; in other words, its acceptability.
– This justice can only be measured through the agreement and consensus achieved through the different participation mechanisms allowing free interaction and which maintains its critical awareness in the public sphere (Habermas 2022, 100).

The democratic legitimation of political action can never be based only on an electoral act. Instead it requires a stable, trustworthy connection into a public process of debate and decision-making without which we cannot speak of a common will. In this way, democracy can be understood as a system supported, guided and even controlled by opinion generated in this public sphere. From these basic functions it is possible to agree an approach to the concept. As Sartori states, "Opinion is called public not only because it is from the public (spread among many or among the majority) but also because it affects objects and subjects that are public in nature: the general interest, the common good, the *res publica"* (1992, 149). Although this definition is close to the intuitive idea we have as citizens, there is a third dimension: space. Three basic dimensions are interlinked, then, in this concept of a public sphere and its relationship with democracy (Rabotnikof 2002).

Firstly, we are talking about *opinión*, or *doxa*; fallible knowledge by definition open to revision that has no horizon of reference beyond agreement among those affected. As we will see in the next point, it is not a question of reducing reality, their needs and interest, the values from which we build it, to data made up of algorithms that give them meaning and utility, by the way, based on scientificism and technical interest of very clear dominion (Habermas 1984). They are opinions, not facts; reasons, not motives. Positions taken before a political and social reality that never is or can be independent of the very subjects making it up. Cognitive content formed via interaction with flows of information about public affairs requiring being disposed to debate. The result of this participation consists, as Habermas says of "(…) communication flows being filtered and summarized in such a way that they are condensed into public opinions around specific issues" (Habermas 1992: 440).

Secondly, the word *public* is being used to refer to the set of citizens

involved in and/or affected by a political decision—or, as appropriate, their representatives—and to refer to the type of content affecting everyone equally. So, by *public* we not only understand the object, but also the subject. In other words, an opinion is public when it has two characteristics: spread among publics and reference to public affairs. However, not all opinions reaching the media or parliamentary mechanisms and that represents the opinion of a group deserves to be called just. As we will see, it is necessary to introduce the quality of public opinion and not the quantity if we want to explain its influence.

Thirdly and finally, we are talking about a *public* space; an inclusive, open sphere of participation where only equal conditions of participation can ensure the generation of public opinion. On one hand, civil and political rights guarantee the formation of an open, plural space: legal equality between citizens, freedom of thought, expression, assembly, press and association, the possibility of political participation and so on. These guarantees have precisely the basic function of allowing the free flow of public opinion to our representatives and, in general, of constructing a public sphere as a set of the different forums in which citizens appear as the public. In summary, a space where participation, debates and deliberation produce quality (or autonomous) public opinion. It is a space belonging to civil society where the exchange of opinions takes place (Sennett 2011).

Based on this triple dimension, if the agreement reached is the result of reason, of debated opinions, it would be linked to an *autonomous or true public sphere*. If, on the other hand, the opinion formed is the result of deceit, manipulation, persuasion, pressure or influence, it would be a heteronomous or spurious public opinion. We can add whatever predicates we like, but in this second case we are not talking about democracy. More recently, Habermas sets out a series of preconditions for being able to talk about an autonomous public sphere and failure to meet these conditions has a great deal to do with the current deterioration of deliberative forming of opinion and the common will (Habermas 2022, 30).

Firstly, citizens require a liberal political culture consisting of a delicate network, as Habermas tells us, of attitudes and cultural assumptions reflected in constitutional principles. It is a culture that requires orientation towards the common good, a form of civic solidarity which, far from demanding individual altruism rather involve “reciprocal willingness to help”.

A second necessary condition for an active civil society, as a social basis for the public sphere, is a level of social equality making it possible for the electorate to participate equally, thereby guaranteeing the political autonomy of citizens via rights of communication and public participation; ultimately, an

inclusive public sphere. Inequality leads to exclusion and this, as we have seen in populism, to mistrust and punishment votes.

This condition of equality and inclusion was already present in *Faktizität und Geltung.* Habermas puts it clearly: freedom of thought, freedom to freely access all sources of thought, also requires social and material conditions that guarantee in principle, equal capacity to participate (Habermas 1998, 456). “Merely formal recognition” of rights is not enough to ensure the formation of public opinion. Equality must go beyond abstract and formal dimensions and achieve the necessary minimums, primary goods, we might put it, so that a person can, if they wish, participate in determining opinion and the common will.

Thirdly, and directly related to inequality, is the relationship between the democratic state and the capitalist economy. As we well know, the tension between a necessary valuation of capital in order to generate fiscal income and the necessary redistribution to ensure the legal and material conditions for the autonomy of each citizen. If this is not the case, as we have seen in Chapter One, states lose credibility and democratic legitimacy (Habermas 2021: 483). For Habermas, only the harmony of the public sphere based on civil society with the formation of opinion and will legally institutionalized in parliament offers sufficient explanation of what, as participants and citizens, we assume democracy is.

This relationship between democratic institutions and an inclusive sphere of participation and possible acceptance today legitimize our social and democratic states under the rule of law. However, as we have seen, these conditions are not being met and this bad situation has led to a *new democratic revisionism* where democracy is no longer put forward as the only option in the face of inequality and the corresponding disaffection (García-Marzá 2015). The mistrust of the pillars supporting the democratic system also reach the concept of the public sphere. The key question we must answer is whether today public opinion is an autonomous opinion or whether instead we must forget about the concept based on structural change. To refer to this change, we will use the concept of *artificial public opinion*.

5.2 Platformisation: A New Structural Change in the Public Sphere?

In his work Facticity and Validity (2022), Habermas recovers the genuine meaning of the concept of the public sphere (Öffentlichkeit) and delves into the process of its formation. What defines the autonomy of public opinion is not its

representative character, the quantity of opinions collected, but its quality, and this depends rather on the way in which it has been produced, that is, on the process of formation itself. In this sense, Habermas speaks of the public sphere as a “network for the communication of contents and positions, i.e. opinions, which is reproduced through communicative action”. Its validity would be determined by the same rules that determine the validity of communicative action, that is, by the distance that separates or unites them with the presuppositions of a dialogue free of any external or internal pressure, where equal conditions of participation are assured. In short, their quality is determined by the procedural properties of their formation process. The relevance and effectiveness of the public sphere depend on this quality. That is, its capacity to influence, its power, in short. According to Habermas, although public opinion is generated “in face-to-face encounters in everyday life and in public events, public communication is the only domain in which the noise of voices can be condensed into relevant and effective public opinions” (Habermas 2021, 485).

In fact, Habermas clearly differentiates between two types of actors who move through the public sphere. On the one hand, there are the autochthonous authors, who reproduce the public sphere itself through their actions, based on understanding and possible agreement. On the other hand, there are the usufructuary publics, those who use this sphere to achieve their particular interests, to instrumentalise the public. Among the latter are political parties, interest associations and the mass media. Habermas’ proposal is that the same conditions of transparency, publicity and participation should be imposed on them as were once imposed on the political sphere. The question is now, 30 years later, whether it is possible for the media to remain the infrastructure of the public sphere, or whether the current process of digitalisation brings with it new dangers that the media cannot cope with. Who are now the guardians of the communicative currents that are generated in the public sphere? Can moral and political autonomy be the object of usufruct?

Today, this situation has changed, but, so to speak, for the worse (Seeliger and Sevignani 2021). The digital revolution in which our society and, therefore, our democratic systems find themselves, has increased these problems, adding a second disruptive element to the hitherto used notion of public opinion. Now the information published and the design, analysis and interpretation of polls are intrinsically linked to artificially intelligent mathematical models. It is no longer the media that are responsible for truthfully, truthfully and fairly conveying the public opinion generated in the public sphere. Now it is the technological platforms that feed on the massive data and metadata generated by social network accounts, which are largely managed by social bots; that is, by Artificial

Intelligence algorithms designed to generate content and/or maintain interrelationships in cyberspace. On the other hand, the collection and interpretation of the massive data and metadata generated by these accounts is carried out by other types of Artificial Intelligence algorithms, those designed specifically for this type of task (Harper 2017, 428). This artificial public opinion, fabricated from the technical interest of domination, is a basic dimension of what we have been calling algorithmic democracy, a framework of meaning and action, whose features can be synthesised as follows (García-Marzá and Calvo 2022)

1. The ideology of neutrality. Although we will deal later with the critique of algorithmic reason, the belief that algorithms are objective and neutral is the frame of meaning of this progressive dissolution of the public sphere. There is no need for deliberation about the common good. What we really want, even the calculation of the consequences of our decisions, can be translated into data and organised by algorithms. Participation and dialogue are not necessary, rather counterproductive, the democratic experience of the average voter is enough. It is a new form of opinion representation and manufactured from the generation, collection and exploitation of massive data and metadata on opinions, behaviours and experience of the citizenry. The aim of all this is to obtain an average representation of the electorate that serves as a reference for managing public affairs, to foresee possible anomalies and conflicts in democratic processes and to control both the system and the citizenry. As Christina Larson says: “who needs democracy when you have data” (Larson 2018).
2. Platformisation: the reduction of the human being to a mere terminal for the flow of data and metadata online. The citizen thus becomes the main product of the emergence of the subsistence and development of data superpowers, of so-called attention capitalism. The value of human beings in politics is established on the basis of their ability to feed artificially intelligent mathematical models that design or help design public policy, make or help make state legislation, and make or help make political decisions, among other things. In this field of distortion of social reality, the Metaverse is the last rung in the process of displacement of public space towards hyperconnected, anomic and commercialised virtual environments where human beings offer algorithms privileged access to their private and, above all, intimate world. The networks and the platforms that support them are responsible for converting the public sphere into a plurality of semi-

public spheres where it is now difficult to separate the public from the private, the commercial from the political, the individual from the singular. In short, where there are no longer the pretensions of validity, nor the possibility of arguing them, that structure all communicative action.

3. The pre-eminence of online data and metadata in the management of public affairs. The increased effectiveness of public policies, laws and political decisions are seen as the fruit of online data and metadata that are consciously and unconsciously generated by digitally hyper-connected citizens (Caplan and Boyd 2016; van Dyk 2021). There is, therefore, no interpretation and reflexivity on the particular, collective and general expectations underlying democratic practice, only a process of computational calculation of majorities on different elements and decisions of the system. A process in which designers and developers of algorithms are both sufficient and surplus to requirements. The participants and those potentially affected are neither here nor expected (Peirano 2019).

4. Surveillance democracy: design and use of ecosystems for mass surveillance of citizens. Intelligent mathematical models are designed, applied and used in different contexts of interaction to track, collect, store and process the huge and heterogeneous amount of online data and metadata generated by digitally hyper-connected citizens in different democratic processes (van Dijck 2014; Calvo 2019). These surveillance ecosystems recreate a cyber-physical space that, controlled and governed by artificially intelligent algorithms, allows digital hyperconnectivity, mass tracking and behavioural datafication with or without citizens' consent through sensors, cameras, computers and other fixed or mobile technological devices (García-Marzá and Calvo 2022; Calvo 2019). We are facing a democracy in which algorithms are defining the processes of public opinion formation, decisionmaking in public policy, calculation of results, etc. It is not that computers have learned to think and make decisions for themselves, but "we have learned to let them think for us" (Runciman 2019, 151). The citizenry is no longer the target but the product, and public opinion is increasingly manipulated by the massive doxa and artificial echo produced by social and political bots, with or without Artificial Intelligence in the different digital spaces of participation and social interaction. Public opinion is no longer controlled by humans. In fact, the hybridisation between the human and the non-human distorts the boundaries between the real and the virtual and, with it, deforms the processes of public management and their results (Donati

2019a). The autonomy of the will dissolves in the face of the seduction of objectivity, algorithmic paternalism, and the so-called technochovism that invades us (König and Wenzelburger 2022).

The main problem with this conception of democracy, conceived and designed from Artificial Intelligence, is that the deficits of participation implied by the progressive substitution of people by data, the gaps in the formation of opinion and common will, have been rapidly filled and occupied by information, generated by platforms and their algorithms endowed with Artificial Intelligence, which today constitute the decisive impulse in this new transformation of public opinion. What is most worrying is that this supplanting is not merely a confusion between public opinion and common opinion as a massive data tally of individual opinions left by humans and non-humans in cyberspace. What is worse is that this opinion thus shaped merely reflects in polls on massive data and metadata that reality that has previously been constructed by technological corporations that only respond to mercantilist logic, to the systemic imperatives of power and money (Harper 2017). They admit no responsibility, not even professionalism when it comes to disseminating information that they have not produced.

In short, just as an algorithmic democracy cannot exist in name only, neither can we speak of an artificial public sphere. A critical perspective must denounce, from the political autonomy that democracy presupposes, this ideology. In this new structural change, none of the three dimensions of the public sphere are fulfilled: neither inclusion, nor public discussion for the common good, nor a space where free and equal participation is possible. Another issue would be to expand the political public space into civic or civil public space (Brinkmann et al. 2022).

At the time of writing, efforts to regulate these corporations are the order of the day, but the results are different, given the complexity, acceleration and global nature of their activities. The proposal of this book is complementary to this legal regulation. It is about confronting this algorithmic democracy from civil society, designing within the same platforms, spaces for participation, civic public spheres, where the inclusion of all the parties involved (stakeholders) have a voice and where the primacy of communicative action, of understanding and agreement, constitutes the criterion of validity for their quality and autonomy.

5.3 Cyber-physical Ecosystems, Big Data, and Bots:

The Dissolution of the Public Sphere

At the dawn of the twenty-first century, and with the technological convergence of the internet, web 2.0 and digital devices, especially the mass use of social networks such as Facebook and mobile phones, CMC was discussed by Anabel Quan-Haase and Barry Wellman in terms of *hyperconnectivity* and as "the availability of people for communication anywhere and anytime" (Quan-Haase and Wellman 2005, 285). This trend generated a new communication ecosystem that allowed, among other important things, the complete restructuring of social relations and forms of information circulation, and these came to be understood in a completely new way. More importantly, the multiple interactions established by the ecosystem's users gave rise to a continuous flow of data, which could, in turn, be analysed, managed and strategically used in different fields of application, such as economics, advertising or politics.

As José van Dijck argues, as engineers swiftly found ways to encode all that information into algorithms that help form a specific model of online society, connectivity soon evolved into a valuable resource. The largest and most influential platforms —Facebook, Twitter, Google, and others— saw a real explosion in user numbers and their potential to generate money. The interconnection of these platforms give rise to a new infrastructure: a connected media ecosystem. The shift from networked communication to a society shaped by digital platforms, as well as the transition from a participatory culture to a true culture of connectivity, occurred over a brief period of no more than ten years (van Dijck 2013, 2014).

Today this trend has reached its utmost degree of sophistication through what is termed *digital hyperconnectivity*: the potential, and people's propensity for remaining permanently connected to *everything* they consider important —not just people— through a continuous flow of online data and metadata. The IoT has made this development possible, and it has also permitted the creation of *cyber-physical ecosystems* structured around digitally interconnected software, devices, sensors that are controlled virtually by AI-equipped algorithms. This, in turn, has enabled the digital hyperconnectivity of all reality: people, animals, processes, behaviours, elements and things. Now, what keeps us connected are data and metadata, not language. In other words, in the current era of digitization, communication is understood as the exchange of all kinds of data and metadata, which makes it possible to be continually connected with any user or element of the *cyber-physical ecosystem*.

So, in the twenty-first century, the sensory reality of the world is underpinned by an infrastructure with massive volumes of data and metadata.

Today, whoever controls these immense volumes of data and metadata —especially the Big Tech companies (Saura-García 2022)— controls the structure that sustains this system, and with it the diverse—cultural, social, personal—spheres of the lifeworld. In this way, they can orchestrate this increasingly hyper-data-driven, hyperconnected and hyper-ICT-dependent world as they wish. They are the hyperaesthetic *masters of mankind* of the hyperdigital era and have a *quasi-limitless* capacity to transform the natural and sensory world, as well as influence the construction of meaning, truth and justice. However, this power threatens to use systemic imperatives, such as power and money (to quote Habermas again) to colonize the rationalization processes underpinning the meaning that beings endowed with speech and action deploy in order to reach communication-mediated, intersubjective agreements on issues linked to knowledge generation and the coordination of action. Essentially, these non-linguistic media fill the gap and, at least to some extent, replace everyday language (Ritzer 1995, 510–511). Thus, instead of language guiding and coordinating action, this increasingly complex system releases systemic imperatives that exhaust the capacity of the lifeworld they instrumentalize (Habermas 1987: 155). The violence exerted upon the lifeworld by this system's various ways of restricting communication, in turn, leads to the emergence of *pathologies* in the lifeworld. Today the *colonization of the lifeworld* is undertaken through *Big Data* and the application of its practical techniques and technologies.

The term *Big Data* was coined around 2000 in fields such as astronomy and chemistry (Mayer-Schöenberger and Cukier 2013) and used to refer to the speed, variety, versatility, value and volume of online data and metadata. It also referred to the attempt to analyse the major flow of structured and unstructured data and metadata that was steadily being generated by the cyber-physical ecosystems and technological platforms involved in the twenty-first-century citizen's ever-increasing tendency to be constantly communicating via mobile devices interconnected through the internet with whatever they consider to be relevant —people, animals, processes or objects of all kinds— at any time, anywhere. Since then, different disciplines (such as data science), techniques (such as *data mining*, *clustering*, *association roles*, etc.) and technologies (such as large-scale databases or data analytics tools) have been developed around *Big Data*. On one hand, these have been used to collect, store and process massive volumes of data and metadata of all kinds, while, on the other, they have been used to convert all this data and metadata into relevant information and then into applicable knowledge through the use of Artificial intelligence (Calvo 2020).

The consolidation of *Big Data* following the attacks launched on 11

September 2001 became a turning point for modern societies and their different fields of activity, and not always for the better (Zuboff 2018). On one hand, *Big Data* offers major opportunities to improve a range of processes—manufacturing, care, clinical, communication, participation, decision-making, creative, research and innovation—from a range of perspectives, including sustainability, predictability, speed, comprehensiveness, extensibility, capacity, completeness, consistency, efficiency, ratification, accuracy, detection, entertainment and creativity (Marr 2015). On the other hand, there is also a darker side to the use of *Big Data*: its direct or indirect involvement in the exponential increase in social complexity, which entails greater levels of vulnerability, uncertainty, inequality, disaffection, instrumentalization, reification, heteronomy, alienation, anomie or psychopathologies for society and citizenship, above all for society's most vulnerable groups (van Dijck 2014; O'Neal 2016; Calvo 2019, 2020, 2022; González and Calvo 2022; Vayena and Gasser 2016).

The recreation of cyber-physical ecosystems and the Big Data emergency have given rise to another no less important and decisive phenomenon, which we referred to earlier: *datafication*. This term is a neologism also coined by Victor Mayer-Schöenberger and Kenneth Neil Cukier in *Big Data*. *A Revolution that will Transform how we Live, Work, and Think* (2013: 73–97) intended to refer to the processes of transforming social action into quantified online data which enables companies and government agencies to carry out real-time predictive monitoring and the analysis of digital citizenship through AI-enabled algorithms (van Dijck 2014: 198). The high expectations raised by *datafication* have generated a strong trend towards the algorithmization of different areas of human activity, including those linked to democratic processes. In this regard, the fact that control, security and decision-making are increasingly being left in the hands of AI algorithms is particularly significant because of their supposed ability to analyse the massive volumes of data which AI constantly generates; as predict the impact and consequences of any possible decisions that may need to be taken; and act according to a profit-maximizing criterion.

Within the democratic sphere, the potential of *Big Data* and the different digital techniques and technologies used to exploit it, such as *data mining*, large databases or *data analytics* tools, can help develop closer, more effective and efficient public policies through a deeper knowledge of the preferences, interests and opinions of digitally hyperconnected citizens, while also developing a greater and faster capacity to detect hidden behavioural patterns and disruptive anomalies which can help establish sustainable democratic systems. Big data can also be used to: design and apply technologies capable of detecting corruption,

bribery or tax evasion with greater precision, as well as more robust, traceable and predictive public health and security systems; implement swifter, more efficient management processes and procedures for any embedded conflicts by significantly increasing the capacity to predict citizens' expectations and demands; and, finally, improve voter engagement or loyalty through personalized advertising campaigns, among many other things (Napoli 2014). However, as will be discussed below, the expected improvements have increasingly become risks and real dangers for democracy.

One of the first times that *Big Data* was applied and used in politics was the 2008 US presidential election (van Dijck 2014), when Barack Obama's team decided to apply predictive models based on HP Vertica analytical data in order to gain a deeper understanding of the undecided electorate, and thus launch personalized advertising campaigns through social networks to capture their vote.

A year later, in 2009, what was possibly the first attempt to apply *Big Data* techniques and technology to the governance of a country was developed. The Obama administration promoted and funded the design, implementation and use of mass data and metadata analysis systems to be used for managing border control, health homeland security, and national infrastructure. A key step was taken with the implementation of the *Open Government Directive*, which obliged government agencies to disclose their data, which was followed by the creation of the online portal Data.gov, created in 2009 to comply with this *Directive*. In 2012 these were followed by the *Big Data Research and Development Initiative* in 2012 (Kalil 2012), which was assigned a task list with more than 85 actions across different fields and had an annual budget of 250 million dollars per year. Finally, in 2015 the first Chief Data Scientist of the United States was appointed. As Tom Kalil, *Deputy Director for Policy at OSTP*, declared in a White House statement issued on the launch of the *Big Data Research and Development* programme, “By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning” (Kalil 2012).

The results obtained from *Big Data* by this and other governments, which included substantial improvements in political decision-making, public policymaking and fraud reduction, were accompanied by a heady discourse concerning its development and application in different spheres of human activity. Together they have driven the rapid expansion and acceptance of *digital hyperconnectivity* by both the informed public and the citizens in general (Marr 2016). However, this process gives rise to highly corrosive consequences, both

intended and unintended, for the users of the system and society (van Dijck 2013, 2014). In this regard, as mentioned above, it is worth highlighting state agencies and political parties' fraudulent use of *bots* to manipulate public opinion and interfere in political communication for the purpose of obtaining financial contributions, capturing votes, recreating an artificial aura of popularity, interfering in rivals' communication strategy and launching false opinions and hoaxes in order to influence the free will of the electorate. In short, they manufacture artificial public opinion, in the sense that this opinion is external to those who are involved and affected.

For example, Obama was criticized in the US for flooding social media with automated messages in an attempt to attract public attention and support in both the 2008 and 2012 elections; Mitt Romney, a Republican candidate in the 2012 US elections, was accused of buying thousands of Twitter followers in an attempt to appear more popular; and, finally, Donald Trump used *social bots* and fake profiles on Twitter and other social networks to launch favourable opinions about his candidacy, increase his followers, generate an artificial perception of greater popularity, and attack his opponents by launching fake news or subliminally distorting their image, among other things (Bessi and Ferrara 2016; Howard et al. 2018; Molina et al. 2017). In France, during the 2017 presidential election, *bots* linked to the candidates Marine Le Pen and Emmanuel Macron were detected in the days leading up to the election (Ferrara 2017, 2020). In Brazil, Jair Bolsonaro was accused of circulating millions of fake news stories through social networks such as WhatsApp during the 2018 presidential elections and, while in government, he was criticized for spending the equivalent of 1.8 million euros on social media advertising mainly to generate a positive image of his handling of the COVID-19 pandemic. In Mexico, a study showed that during the 2018 presidential elections 53% of the Twitter followers of the presidential candidates of the 2018 elections —Andrés Manuel López Obrador, Ricardo Anaya, José Antonio Meade and Jaime Rodríguez Calderón— were not people, but *social bots* (Barabási and Ruppert 2018). In Spain, politicians multiplied followers and *likes* on Twitter during the 2019 presidential elections and the COVID-19 pandemic, and there was an average of around 42% for each of the country's main political leaders (Lagares et al. 2021; Manfredi et al. 2021). Finally, in the UK, as revealed by *The Guardian*, the pro-Brexit campaigns, led by, among others, the future British Prime Minister, Boris Johnson, hired the services of Cambridge Analytica and AggregateIQ to influence citizens by applying *data mining* techniques to the mass volumes of data and metadata generated by users of social networks such as Facebook; data that was fraudulently obtained (Cadwalladr 2017).

These and other practical examples of the digital transformation of the public sphere and its artificialization using social bots call for a searching reflection on the ethical implications of algorithmic governance, the technological colonization of the public sphere and the possible negative effects of *digital hyperconnectivity* on society. Above all, this is necessary because *digital hyperconnectivity* and everything linked to it is underpinned by information, and this is a public good, not a private one. It is not possible to consume information without it affecting our understanding of the world, as well as our individual will (aspirations, desires, interests, and so on). Likewise, its consumption affects what we might call, following Habermas (2009), the formation of a collective will as a society; in other words, the formation of public opinion, which is the basic pillar of democracy (García-Marzá and Calvo 2022).

5.4 Generative Social Bots: The Construction and Development of the Artificial Public Sphere

In this hyperconnected, datafied, algorithmized and robotized world, the main problem we face is that new digital technologies are amplifying the power of lies, manipulation and deception, destroying the reputation of and our trust in the media and the strength of public opinion.

The process of destroying all references to truth and justice always follows the same path. Firstly, the concept of reality is destroyed. Today, for example, there is talk of deceitful concepts such as post-truth, as if there were something beyond truth; and as if it were possible to coordinate action without an intersubjective aim of attaining truth and justice. This is a scenario where reality is defined in accordance with the wishes of each person, and where those who accumulate more power have more opportunity to persuade and distort reality. That makes justice relative and arbitrary, and there is nothing to guide people's actions; essentially, each person defines their own course of action as they see fit in, accordance with their own interests and desires.

However, trust in the democratic process depends on a series of expectations being met, and if we know we are being manipulated, lied to, or deceived, we stop voting for a political party, buying a newspaper, watching a news channel, participating in political decision-making processes, and so on. The major centres of power are therefore using the new digital technologies to artificially generate credibility and *artificial confidence* by creating distorted fields of reality which are so seductive and absorbing they evade any criticism from the digitally hyperconnected citizenry (García-Marzá and Calvo 2022).

In this regard, it should be noted that over the last decade *social bots* have become one of the main disruptive elements affecting the democratic process. Above all for their intrusion into processes linked to generating and developing public opinion. In the wake of the emergence of the internet and related phenomena such as social networks, digital hyperconnectivity and *datafication*, cyberspace has gradually become the principal forum for the public sphere, replacing the traditional spaces of participation and *relations* (Hagen et al. 2017). Given the high number of artificially intelligent algorithms interacting in cyberspace, and the increasing difficulty in distinguishing between human and non-human interlocutors, especially *social bots*, these developments have led to social relations becoming disfigured and dehumanized, and the results are ever more apparent: the accelerating collapse and demise of public opinion.

Social bots are social network accounts that may, to a greater or lesser extent, be controlled by artificially intelligent algorithms. They are able to interact with people and machines and even generate specific content in order to achieve some kind of strategic goal (Yang et al. 2019; Hajli et al. 2021; Ferrara et al. 2016). Intelligent *social bots*, whether they are political or commercial (and nowadays it is difficult to distinguish them) are not malicious *per se*. They can perform a variety of useful and beneficial functions for agencies, corporations or individuals, such as reducing human labour, solving problems or providing information and advisory services. Nevertheless, these *artificial social actors* can also carry out many harmful actions, such as providing people with misinformation, spreading arguments, perpetrating scams and exploiting the stock market (Ferrara et al. 2016; O'Neal 2016). In other words, just as traditional malware—such as computer viruses, worms, Trojans, ransomware, spyware, adware, scareware, etc.—attack the weak points of hardware and software, *social bots* exploit human vulnerabilities to achieve a given goal, for example, people's inability to distinguish between human and non-human agents when interacting in cyberspace, their overconfidence in online social relationships, the fact that people pay increased attention to everything that seems clouded in an aura of popularity, or simply their simple naivety in the face of the manipulation of public opinion conducted through social networks (Hajli et al. 2021).

In this sense, there are many types of *social bots*. Many of them are fairly simple, cheap and as easy to track as they are to build, but others are sophisticated and adopt various strategies to impersonate humans, which makes it difficult to track them. With regard to the sophisticated ones (Hajli et al. 2021), some models of social bot use state-of-the-art *machine learning* to enable them to learn to communicate like humans and go unnoticed; others adopt names,

identities and profiles of real users to avoid being discovered and continue working towards achieving their given goals; still others temporarily mimic human patterns of content consumption and publication to appear real; and, finally, some even engage in conversations, answer questions and exchange opinions with other users to make themselves more elusive and avoid being discovered.

In the democratic sphere, the main function performed by *social bots*, and their key areas of impact, are: generating popularity, interfering in an opponent's communication and influencing public opinion. These three functions are performed as follows:

– *Generating popularity*: governments, parties and politicians buy a certain number of *social bots* to convert them into their (fake) followers and these bots (falsely) praise them and their policies, actions and decisions to boost their online reputation so that they gain real followers, spread their ideas more widely, and gain support and votes (Cresci et al. 2015, Finger 2015).
– *Influencing public opinion*: governments, parties and politicians use the most sophisticated *social bots*, those equipped with *machine learning*, to interact with other social network users and promote trends, encourage ideas, comment on posts and argue for or against a given topic. We could say that they try, and in many cases succeed, to replace the legitimizing role of public opinion along with its influence on the formation of a common will. Indeed, "(…) alarmingly, the activity of *social bots* accentuates their effectiveness during political election campaigns with their efforts to mislead citizens around the world towards through manipulated viral trends" (Pastor-Galindo et al. 2020).
– *Interfering with competitors' communications*: governments, parties and politicians can also use the most sophisticated *social bots* to interfere with competitors' communication campaigns by circulating *fake news*, exchanging contradictory information, or else through the purchase of poor-quality social *bots* to misrepresent their competitors. As *bots* are easily tracked, they are quickly identified and the competition—which may be a government, a party or a politician—are discredited and labelled as cheats (Woolley and Howard 2018).

Although, these are not the only functions linked to the design, implementation and use of *social bots* in the political sphere, they constitute a significant challenge for the health of digital democracy. Politicians such as Newt Gingrich, Mitt Romney, Barack Obama, Jair Bolsonaro or Donald Trump, political parties such as the conservative Christian Democratic Union of

Germany (CDU) and governments such as China, the US and Russia have been widely reported as using *social bots* to perform these and other functions in order to achieve their political goals and interfere in the operation of other democracies.

Nowadays, the use of social bots in politics has now become a widespread, pervasive and somewhat socially tolerated practice. For example, the study "Spotting political social bots in Twitter: a use case of the 2019 Spanish general election" (Pastor-Galindo et al. 2020) showed that all the main political parties who had participated in that year's Spanish presidential elections used *social bots* to appeal to the electorate, interfere in the communication emitted by the competition and/or generate public opinion favouring their strategic interests; over 800,000 accounts were identified as susceptible of being manipulated by social bots. Likewise, the study "Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection" by the US National Bureau of Economic Research found that Twitter bots directly influenced 3.23% of votes in the 2016 US presidential election and 1.76% of votes in the 2016 Brexit referendum (Gorodnichenko et al. 2021).

Thus, as the study "Bots, elections, and social media: a brief overview" (Ferrara 2020) demonstrates, the activities carried out by *social bots* during some of the most relevant presidential elections in recent years, such as those held in the USA in 2016 and in France in 2017, raise disturbing issues regarding their use, which includes their widespread participation in social networks during the campaigns, the fact that their behaviour is very similar to that of humans when engaging in social interactions, and their capacity to adapt and thereby go unnoticed by the bot detectors put in place by social networks such as Twitter and Facebook. Therefore, these and other studies suggest that there is a black market in *social and political bots,* which poses a serious threat to modern digital democracies.

As a result, *social bots* have sown reasonable doubts about the democratic process and the public opinion responsible for legitimizing it; they have broken down the confidence underpinning the power and influence of public opinion. The digital transformation of democracy is generating a false perception among citizens that public opinion is highly dependent on ICT applications, so, in order to influence public opinion, it is necessary to engage with the underlying phenomena of digital hyperconnectivity, social networks and *datafication* (Hagen et al. 2017; Innerarity and Colomina 2020; Calvo 2019, 2020).

The main problem with the digitally generated public sphere is that it is based on *cyber-physical ecosystems*, which are in turn governed by artificially intelligent mathematical models serving the particular cause of the government,

company or political party that acquires them and puts them into circulation in accordance with systemic imperatives such as power and money. This is what lies behind the concept of *algorithmic democracy* as a horizon for action (a trend, attitude, etc.) and meaning (algorithmic culture) based on the progressive implementation of Artificial intelligence and its technologies across all processes of deliberation, decision-making and institutional design, both in the state and civil society (García-Marzá and Calvo 2022). The result is a space where developers themselves are unable to meet the challenge of designing tools capable of detecting manipulation in order to purify the public opinion processes undertaken in cyberspace. Essentially, the developers are unable to differentiate between public opinion and *artificial public opinion*. One outcome of the current inability to distinguish between the two, is that there has been an exponential increase in citizens' distrust of and disaffection towards modern democracies, which were already badly blighted by inequality and discredit (Innerariti and Colomina 2020; Keane 2018).

The latest episode of this perverse drift currently affecting public opinion features the entry of generative AI. As shown by the report *Freedom on the Net 2023: The Repressive Power of Artificial Intelligence* (Funk et al. 2023), social bots using generative AI, like ChatGPT from OpenAI and Google Bard, have become the epicentre of a new but exponentially deeper, and more accessible, subtle, invasive, disturbing and oppressive, field of distortion of public opinion.

> These campaigns have long been assisted by AI technology for distributing content: automated "bot" accounts on social media have smeared activists into silence and propagated false narratives about electoral fraud to voters, while platform algorithms have promoted untrustworthy and incendiary information over reliable sources (Funk et al. 2023).

The report reveals that governments all over the world, with all kinds of ideologies, have begun to use generative AI in the digital sphere for despotic, instrumental and malicious purposes. With this, governments intend to achieve competitive benefits and increase their power and control over society, strangling spaces for freedom, undermining the public sphere, deconstructing the available information at will, manipulating their own or others' electoral processes through processes of disinformation, and producing artificial mass public opinion to bring meaning and legitimacy to their authority, among other things.

Unlike previous models of social bot, generative AI has high capacity for natural language, which allows it to hold realistic conversations with human

beings continuously, progressively and elusively. In addition, however, these social bots can use other types of communication channel to achieve their ends. "(…) making the truth easier to distort and harder to discern" (Funk et al. 2023, 2). For example, using mass images, videos and voices, both from fake profiles and ultrarealistic emulations of influencers from communication, society and politics (deepfakes) to construct a convincing, aggressive and highly addictive and influential story capable of capturing attention and altering public perceptions.

> AI-based tools that can generate text, audio, and imagery have quickly grown more sophisticated, accessible, and easy to use, spurring a concerning escalation of these disinformation tactics. Over the past year, the new technology was utilized in at least 16 countries to sow doubt, smear opponents, or influence public debate (Funk et al. 2023, 1).

In constructing this new, disruptive truth distortion field, the private sector has become a great ally of governments, political parties and candidates. In Russia, for example, different technology companies providing digital products and services have played an important role in the disinformation campaign orchestrated by the Kremlin on the invasion of Ukraine, such as Operation Doppelgänger, where they "mimicked German, American, Italian, British, and French media outlets to disseminate false and conspiratorial narratives about European sanctions and Ukrainian refugees, among other topics" (Funk et al. 2023, 1). To make this more effective, Kremlin approved aid to the private sector to develop its own generative AI (President of Russian 2023). Israel, which also has an incipient but growing market of paid disinformation companies, uses AI to create false narratives based on concepts and keywords and online digital platforms to mobilize a network of false accounts on social networks to spread and promote them. Particularly important is the proliferation of disinformation and news manipulation liked to the war between Israel and Hamas in the Gaza Strip, in which the use of generative AI by both sides has shown the disruptive potential of this new technology for public opinion (Hsu and Thompson 2023).

Thus, Generative AI is flooding the public sphere with synthetic datasets whose high level of sophistication, complexity and realism makes them potential and dangerous distorters, deformers and disfigurers of public opinion. In the public sphere, such synthetic datasets are mixed or recombined with real datasets in order to go unnoticed and enhance their results: to introduce the necessary distortion and noise to legitimise or delegitimise the activity and impacts of all kinds of bodies (governments, business corporations, etc.). It is, therefore, a new

phenomenon of digitalisation, that of the synthesisation of reality by artificial means, which, in the case of the public sphere, leads to the massive generation of synthetic opinion and the recreation of an *algosphere*.

In conclusion, within this *algorithmic democracy*, the great danger which public opinion must counter is no longer the tyranny of the majority—as Tocqueville (1989) claimed— but the *tyranny of algorithms*, known as algocracy (Aneesh 2006, 2009; Danaher 2020; Marrone 2022; Gritsenko and Wood 2022), with its apparent objectivity and false neutrality, as well as its assimilation of what is deemed correct from the data and metadata collected, processed and reused by *social bots* and other artificially intelligent technologies. In the absence of a critical perspective capable of explicitly foregrounding this ideology and the false beliefs it is based upon, public opinion and democracy, as we understand them today, will become meaningless. This development is already foreshadowed by the use of expressions such as illiberal democracy and elected dictatorship.

However, democracy cannot be reduced to an unwilled alien aggregate of predefined and unchangeable interests, for the simple reason that this is to disregard our autonomy and our capacity to deliberate, to become convinced and to transform what concerns us. The latter capacity, expressed in our everyday moral language, allows us to discuss misguided interests and undertake a reasoned critique of decisions made by our representatives, even if they are the majority. As is well known, majority does not imply truth, let alone justice (García-Marzá 2015).

In this regard, a deeply worrying development is that governments, major economic corporations and political parties are designing, applying and using artificially intelligent mathematical models on technology platforms, especially *social bots*, to manufacture *artificial and synthetic public opinion* "at will". As Pierpaolo Donati (2019a, b, 2020, 2021, 2023, 2024) has warned, these *social bots* are generating a hybrid form of human and non-human within the sphere of social relations, and their most important consequence is an increasing mistrust in both public opinion and the power it wields.

In the digitally hyperconnected society, interpersonal relationships increasingly take place in cyberspace where people are no longer able to discern whether they are interacting with other people or AI algorithms. As a result, it is possible to interfere in the symbolic production of the lifeworld [Lebenswelt] and its culturally transmitted and linguistically organized stock of patterns of interpretation, which constitutes a repository of self-evidence or unquestioned convictions. This is used by the participants in communication in cooperative interpretation processes to endow with meaning the range of things that make up

the objective, social and personal world (Habermas 1987, 176). The arguments that give meaning to and legitimize political decisions, which were previously in the hands of civil society as the designated space for public opinion, are today subject to the enormous quantity of mathematical models that the system reproduces and applies. Drawing on the wealth of opinions these models produce and the social interactions they cause in cyberspace, they muster and create the reality that interests them. In other words, this situation is wholly opposed to the basic pillar of democracy, public opinion, which is based on the free and equal participation of all parties involved and affected in the democratic system.

Nowadays, moreover, this is coupled with the mass production of synthetic public opinion through governmental and corporate promotion of an *artificial public sphere* or *algosphere* -when the public deliberative space is so colonised by intelligent mathematical models that human influence is reduced to rates equal or close to zero- and the application of Generative AI models, techniques and technology such as ChatGPT, Bard or the different forms of GAN. Thus, if previously the main concern of governments and corporations was how to control, displace and/or minimise the free and reasoned participation of society in the construction of meaning about different aspects of democracy, especially with regard to the legitimisation of the actions and decisions of States and Big Tech, nowadays the concern is how to construct at will, and in a deep, subtle and veiled way, a synthetic reality, sphere and opinion around politics that allows them to increase their control and despotism over society.

Bibliography

Andrino, B., and M. Hidalgo. 2020. ¿Bot o humano? Los políticos multiplican sus seguidores en Twitter durante el estado de alarma [Bot or Human? Politicians Multiply their Followers on Twitter During the State of Alarm]. *El País*. Available at https://elpais.com/tecnologia/2020-04-24/bot-o-humano-los-politicos-multiplican-sus-seguidores-recien-llegados-a-twitter-durante-el-estado-de-alarma.html.

Aneesh, A. 2006. *Virtual Migration: The Programming of Globalization*. Durham: Duke University Press.

———. 2009. Global Labor: Algocratic Modes of Organization. *Sociological Theory* 27 (4): 347–370.

Barabási, A.-L., and P. Ruppert. 2018. Bots and Their Influence During the Mexican Presidential Election – A Network Science Perspective. *Botometer*. http://maven7us.com/bot-analysis/.

Bessi, A., and E. Ferrara. 2016. Social Bots Distort the 2016 U.S. Presidential Election Online Discussion. *First Monday* 21 (11): 1–14. Available at https://doi.org/10.5210/fm.v21i11.7090.

Brinkmann, U., H. Heiland, and M. Seeliger. 2022. Corporate Public Spheres Between Refeudalization and Revitalization. *Theory, Culture & Society* 39 (4): 75–90.

Cadwalladr, C. 2017. The Great British Brexit Robbery: How Our Democracy was Hijacked. *The Guardian*. Available at https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy.

Calvo, P. 2019. Democracia algorítmica: Consideraciones éticas sobre la *dataficiación* de la esfera pública. *Revista del CLAD Reforma y Democracia* 74: 5–30.

———. 2020. The Ethics of Smart City (EoSC): Moral Implications of Hyperconnectivity, Algorithmization and the Datafication of Urban Digital Society. *Ethics and Information Technology* 22: 141–149. https://doi.org/10.1007/s10676-019-09523-0.
[Crossref]

———. 2022. Gemelos digitales y Democracia. *CLAD. Reforma y Democracia. Revista de Éticas Aplicadas* (83).

Caplan, R., and D. Boyd. 2016. Who Controls the Public Sphere in an Era of Algorithms. Mediation, Automation, Power. *Data & Society*: 1–19. Available at: https://datasociety.net/pubs/ap/MediationAutomationPower_2016.pdf.

Cortina, A. 1998. *Hasta un pueblo de demonios: ética pública y sociedad*. Madrid: Taurus.

Cresci, S., R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi. 2015. Fame for Sale: Efficient Detection of Fake Twitter Followers. *Decision Support Systems* 80: 56–71.

Danaher, J. 2020. Freedom in an Age of Algocracy. In *The Oxford Handbook of Philosophy of Technology*, ed. J. Danaher. Oxford: Oxford University Press.

de Tocqueville, Alexis. 1989 [1840]. *Democracy in America, Volume 2*. New York: Vintage Books.

Donati, Pierpaolo. 2019a. *Sociología relacional de lo humano*. Barañain: EUNSA.

———. 2019b. The Digital Matrix and the Hybridization of Society. In *Post-Human Institutions and Organisations: Confronting the Matrix*, ed. Ismael Al-Amoudi and Emmanuel Lazega, 67–92. Abingdon: Routledge.

———. 2020. How to Promote the Dignity of Work in the Face of its Hybridization in the Digital Economy. In *Dignity and the Future of Work in the Age of 4th Industrial Revolution*, ed. S. Zamagni and M. Sánchez Sorondo, 57–90. Libreria Editrice Vaticana: Vatican City.

———. 2021. Impact of AI/Robotics on Human Relations: Co-evolution Through Hybridisation. In *Robotics, AI, and Humanity*, ed. J.S. von Braun, M. Archer, G.M. Reichberg, and M. Sánchez Sorondo. Cham: Springer. https://doi.org/10.1007/978-3-030-54173-6_18.
[Crossref]

———. 2024. El cuidado del otro como relación social. *Recerca. Revista De Pensament I Anàlisi* 29 (1): 1–21. https://doi.org/10.6035/recerca.7527.
[Crossref]

Ferrara, E. 2017. Disinformation and Social Bot Operations in the Run up to the 2017 French Presidential Election. *First Monday* 22 (8) Available at https://doi.org/10.5210/fm.v22i8.8005.

———. 2020. Bots, Elections, and Social Media: A Brief Overview. In *Disinformation, Misinformation,*

and Fake News in Social Media, Lecture Notes in Social Networks, ed. K. Shu, S. Wang, D. Lee, and H. Liu. Cham: Springer. https://doi.org/10.1007/978-3-030-42699-6_6.
[Crossref]

Ferrara, E., O. Varol, C. Davis, F. Menczer, and A. Flammini. 2016. El auge de los bots sociales. *Comunicaciones de la ACM* 59 (7): 96–104.

Finger, L. 2015. Do Evil – The Business Of Social Media Bots. *Forbes* https://www.forbes.com/sites/lutzfinger/2015/02/17/do-evil-the-business-of-social-media-bots/.

Funk, Allie, Adrian Shahbaz, and Kian Vesteinsson. 2023. *Freedom on the Net 2023: The Repressive Power of Artificial Intelligence.* Freedom House. https://freedomhouse.org/sites/default/files/2023-11/FOTN2023Final.pdf.

García-Marzá, D. 2012. Kant's Principle of Publicity. *Kant-Studien. Philosophische Zeitschrift der Kant-Gesellsschaft* 103: 96–113.

———. 2015. El valor democrático de la sociedad civil: una respuesta a la desafección. *Themata. Revista de Filosofiía* (52), 93–109. https://doi.org/10.12795/themata.2015.i52.05.

———. 2016. Enfoques mesodeliberativos: sobre la articulación institucional en las democracias deliberativas. *Isegoría* 54: 147–170.

García-Marzá, D., and P. Calvo. 2022. Democracia algorítmica: ¿un nuevo cambio estructural de la opinión pública? *Isegoría* 67: e17. https://doi.org/10.3989/isegoria.2022.67.17.
[Crossref]

González, Elsa, and Patri Calvo. 2022. Ethically Governing Artificial Intelligence in the Field of Scientific Research and Innovation. *Heliyon* 8 (2): 1–12.

Gorodnichenko, Y., T. Pham, and O. Talavera. 2021. Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection. *European Economic Review* 136: 103772.

Gritsenko, D., and M. Wood. 2022. Algorithmic Governance: A Mode of Governance Approach. *Regulation & Governance* 16 (1): 45–62.

Habermas, J. 1984. *Ciencia y técnica como ideología*. Madrid: Tecnos.

———. 1986. *Historia y crítica de la opinión pública*. Mexico: G. Gili.

———. 1987. *Teoría de la Acción comunicativa*. Vol. II. Madrid: Taurus.

———. 1998. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Cambridge: The MIT Press.

———. 2009. *¡Ay, Europa!* Madrid: Trotta.

———. 2021. Überlegungen und Hypothesen zu einem erneuten Strukturwandel der politischen Öffentlichkeit. In *Ein neuer Strukturwandel der Öffentlichkeit?* ed. M. Seelinger and S. Sevignani, 470–500. Baden-Baden: Nomos.

———. 2022. *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Frankfurt:

Suhrkamp Verlag.

Hagen, L.M., M. Wieland, and A.M. In der Au. 2017. Algorithmischer Strukturwandel der Öffentlichkeit: Wie die automatische Selektion im Social Web die politische Kommunikation verändert und welche Gefahren dies birgt. *Medienjournal-Zeitschrift für Medien- und Kommunkiationsforschung* 41 (2): 127–143.

Hajli, N., U. Saeed, M. Tajvidi, and F. Shirazi. 2021. Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence. *British Journal of Management*. https://doi.org/10.1111/1467-.

Harper, T. 2017. The Big Data Public and Its Problems: Big Data and the Structural Transformation of the Public Sphere. *New Media & Society* 19 (9): 1424–1439.

Honneth, Axel. 2009. *Crítica del poder*. Madrid: Machado Books.

Howard, P.N., S. Woolley, and R. Calo. 2018. Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration. *Journal of Information Technology and Politics* 15 (2): 81–93. https://doi.org/10.1080/19331681.2018.1448735.
[Crossref]

Hsu, Tiffany, and Stuart Thompson. 2023. A.I. Muddies Israel-Hamas War in Unexpected Way. *New York Times*. https://www.nytimes.com/2023/10/28/business/media/ai-muddies-israel-hamaswar-in-unexpected-way.html.

Innerariti, D., and C. Colomina. 2020. La verdad en las democracias algorítmicas. *Revista CIDOB d'afers internacionals* 124: 11–23. https://doi.org/10.24241/rcai.2020.124.1.11.
[Crossref]

Kalil, Tom. 2012. Big Data is a Big Deal [Watch the Live Webcast Today at 2pm ET of the Big Data Research and Development Event at http://live.science360.gov/bigdata/. Available at: https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal.

Keane, J. 2008. *Global Civil Society*. Cambridge: Cambridge University Press.

———. 2009. *The Life and Death of Democracy*. New York: Simon and Schuster.

———. 2018. *Vida y muerte de la democracia*. *Fondo de Cultura Económica*.

König, P.D., and G. Wenzelburger. 2022. Between Technochauvinism and Human-Centrism: Can Algorithms Improve Decision-Making in Democratic Politics? *European Political Science* 21: 132–149. https://doi.org/10.1057/s41304-020-00298-3.
[Crossref]

Lagares, N., P.C. López, E. Jaráiz, and J.M. Rivera. 2021. La comunidad digital y el uso político de las redes sociales: elecciones generales del año 2019 en España. *Araucaria* 23 (48): 235–273. https://doi.org/10.12795/araucaria.2021.i48.11.
[Crossref]

Larson, C. 2018. Who Needs Democracy When You Have Data? *MIT Technology Review* 121 (5): 50–55.

Manfredi, J., A. Amado, and S. Waisbord. 2021. Presidential Twitter in the Face of COVID-19: Between

Populism and Pop Politics. *Communicate* 66: 83–94. https://doi.org/10.3916/C66-2021-07. [Crossref]

Marr, Bernard. 2015. *Big Data – Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. Hoboken: Wiley.

———. 2016. *Big Data in Practice. How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Hoboken: Wiley. https://doi.org/10.1002/9781119278825. [Crossref]

Marrone, P. 2022. Epistemic Democracy and Technopolitics: Four Models of Deliberation. *International Journal of Technoethics (IJT)* 13 (1): 1–14.

Marshall, T.H., and T. Bottomore. 1998. *Ciudadanía y clases sociales*. Madrid: Byblos.

Masuda, Yoneji. 1981. *The Information Society as Post- Industrial Society*. Washington: World Future Society.

Mayer-Schöenberger, Viktor, and Kenneth Cukier. 2013. *Big Data. A Revolution that Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

Minc, A. 1995. *La borrachera democrática. El nuevo poder de la opinión pública*. Madrid: Today's Topics.

Molina, M.D., S.Sh. Sundar, Th. Le, and D. Lee. 2017. Fake News Is Not Simply False Information: A Concept Explanation and Taxonomy of Online Content. *American Behavioral Scientist* 65 (2): 180–212. Available at https://doi.org/10.1177/0002764219878224.

Napoli, P.M. 2014. Automated Media: An Institutional Theory Perspective on Algorithmic Media Production and Consumption. *Communication Theory* 24 (3): 340–360.

O'Neal, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

Offe, C., and U. Preub. 1990. Instituciones democráticas y recursos morales. *Isegoria* 2: 45–74.

Oxford Learner's Dictionaries. 2022. Cyberspace. Oxford Learner's Dictionaries. https://www.oxfordlearnersdictionaries.com/definition/english/cyberspace?q=cyberspace. Accessed 16 Nov 2022.

Pastor-Galindo, J., M. Zago, P. Nespoli, S. López, A. Huertas, M. Gil, J.A. Ruiperez-Valiente, G. Martínez, and F. Gómez. 2020. Detección de bots sociales políticos en Twitter: Un caso de uso en las elecciones generales españolas de 2019. *ArXiv*, arXiv:2004.00931v2 [cs.SI].

Peirano, M. 2019. *El enemigo conoce el sistema: Manipulación de ideas, personas e influencias después de la economía de la atención*. Madrid: Debate.

Pérez-Díaz, V. 1997. *La esfera pública y la sociedad civil*. Madrid: Taurus.

President of Russian. 2023. *Instructions following the Artificial Intelligence Journey conference*. Moscou: Kremlin. http://en.kremlin.ru/acts/assignments/orders/73282.

Quan-Haase, Anabel, and Barry Wellman. 2005. How Computer Mediated Hyperconnectivity and Local Virtuality Foster Social Networks of Information and Coordination in a Community of Practice. In *XXV*

International Sunbelt Social Network Conference. California: ISSNCH.

Rabotnikof, Nora. 2002. El espacio público, Caracterizaciones teóricas y expectativas políticas. In *Filosofía política I. Ideas políticas y movimientos sociales*, ed. F. Quesada, 135–151. Madrid, Trotta.

Ramos, R. 2021. The Covid-19 Crisis as a Political Opportunity. Santiago Abascal's Use of Twitter During the State of Alarm in Spain. *Doxa Comunicación* 33: 75–98.

Ritzer, George. 1995. *Teoría sociológica contemporánea*. Madrid: McGraw-Hill.

Runciman, D. 2019. *Así termina la democracia*. Barcelona: Paidós.

Sartori, G. 1992. *Elementos de teoría política*. Madrid: Alliance.

———. 1996. En defensa de la representación política. *Claves de Razón Práctica* 91: 2–6.

Saura García, Carlos. 2022. El lado oscuro de las GAFAM: monopolización de los datos y pérdida de privacidad. *Veritas. Revista de filosofía y Teología* 52: 9–27. https://doi.org/10.4067/S0718-92732022000200009.
[Crossref]

Scola, Nancy. 2013. Obama, the 'Big Data' President. *The Washington Post*. 14 June 2013. Available at: https://www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html.

Seeliger, M., and S. Sevignani. 2021. A New Structural Transformation of the Public Sphere? An Introduction. *Theory, Culture & Society* 39 (4): 3–16.

Sennett, R. 2011. *The Fall of Public Man*. Nueva York: Alfred A. Knopf.

van Dijck, Joseph. 2013. *The Culture of Connectivity: A Critical History of Social Media*. New York: Oxford University Press.

———. 2014. Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology. *Surveillance & Society* 12 (2): 197–208.

van Dyk, S. 2021, August. Die Krise der Faktizität und die Zukunft der Demokratie. Strukturwandel der Öffentlichkeit in Zeiten von Fake News, Technokratie und Wahrheitskritik. In *Ein neuer Strukturwandel der Öffentlichkeit?* 68–90. Nomos Verlagsgesellschaft mbH & Co. KG.

Vayena, Effy, and Urs Gasser. 2016. Strictly Bioomedical? Stretching the Ethics of the Big Data Ecosystem in Biomedicine. In *The Ethics of Biomedical Big Data*, ed. Brent D. Mittelstadt and Luciano Floridi, 17–40. Cham: Springer.

Velasco Arroyo, J.C. 2006. Deliberation and the Quality of Democracy. *Keys to Practical Reason* 167: 36–43.

Woolley, Samuel C., and Philip N. Howard. 2018. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford: Oxford Scholarship Online. https://doi.org/10.1093/bear/9780190931407.001.0001.
[Crossref]

Yang, K.-C., O. Varol, C.A. Davis, E. Ferrara, A. Flammini, and F. Menczer. 2019. Arming the Public with Artificial Intelligence to Counter Social Bots. *Human Behavior and Emerging Technologies* 1 (1): 48–61. https://doi.org/10.1002/hbe2.
[Crossref]

Zuboff, S. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: PublicAffairs.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_6

6. Moral Learning by Algorithms: The Possibility of Developing Morally Intelligent Technology

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

My thermostat has three beliefs: it believes it's too hot in here; it's too cold in here; and it's just right in here.
John McCarthy to John Searle (Searle 1984, 30)

Abstract
The aim of this chapter is to explore the possibility of moral machine learning and its possible alternatives. We begin by studying artificial neural networks in order to learn about their potential and shortcomings with regard to the creation of artificial moral intelligence. We then undertake a critical analysis of the main AI-based approaches to moral machine learning which use moral value acquisition and encoding processes. The third theme to be addressed is a critical reflection on the possibility of artificial moral intelligence. Finally, we turn to consider whether the state of the art of current AI-based ethical governance offers a plausible alternative to artificial moral intelligence.

Keywords Artificial moral intelligence – Moral machine – Moral machine learning models – Technological singularity – Ethical governance – Self-regulatory

6.1 The Morality of the Thermostat

In one of the Reith Lectures John Searle gave in 1984, published in full in *Mind, Brain, and Behaviour* (Searle 1988), he recounted a conversation he had with

John McCarthy, who had coined the term Artificial intelligence and fostered the emergence of this scientific discipline, following a lecture given by the latter some years before. In his talk, McCarthy not only answered yes to the question posed by Alan Turing a few decades earlier: can machines think? He also went on to state emphatically that any machine, even one as simple as his thermostat, was capable of appropriating "self-beliefs". Faced with McCarthy's bold and unsettling assertions, Searle went on to ask him in a somewhat sarcastic tone about what his thermostat's *beliefs* were, to which McCarthy replied that his "(...) thermostat has three beliefs—it believes it's too hot in here; it's too cold in here; and it's just right in here" (Searle 1984: 30).

Beyond the anecdotal and engaging nature of the example referred to by Searle, McCarthy's words allow us to glimpse how, and the extent to which, the discipline of Artificial Intelligence has, from its beginnings, established a process of appropriating and assimilating extra-technological—biological, social and moral—terminology to recreate a symbolic imaginary capable of anointing, or tokenizing, machines endowed with mathematical models with a kind of naturalization and humanization, or even supra-humanization. In addition to the reconceptualization of the term intelligence for the purpose of defining both the computational capacity of machines and the scientific discipline of Artificial Intelligence itself, the redefinition and linking of terms intrinsically related to human life —belief, learning, interpretation, agency, autonomy, freedom, mind, conscience, thought, judgement, value, justice, neural networks or ethical governance—for the purpose of defining machines' behavioural competence has led to this symbolic imaginary being formed in such a way that artificially intelligent mathematical models are compared to people, or even considered to surpass them both physically and morally. Thanks to this symbolic recreation, machines have become an *inevitable* and *necessary* element for the survival and development of all systems of living beings and human values: given human beings' limitations when it comes to being objective about their reasoning and overcoming the emotional bias in decision-making processes, it is only machines who are able to guide and direct the course of humanity in a just and germane way (Matsumoto 2018).

In this respect, the latest episode in the long process of symbolic construction around machines is to be found in the exponentially greater concern shown by Artificial intelligence with regard to moral machine learning. The development and deployment of driverless vehicles, now dubbed *autonomous cars*, has been slowed down and even halted because of the difficulties designers of artificially intelligent algorithms face when deciding on integrating an appropriate rationality criterion to address or overcome the value conflicts raised by the

practice of *intelligent mobility*. Among other issues, questions can be asked about traffic accident scenarios when it might have to make decisions over whether pedestrians lives would take precedence over those of the vehicles' passengers, or vice versa, which might involve discrimination on the basis of age, sex, race or appearance. There is also the issue of whether animals deserve the same respect and concern as human beings.

For example, *Moral Machine* (Awad et al. 2018) is an online platform whose main aim is to obtain and use online data and metadata concerning the decisions people make around the world when faced with hypothetical moral dilemmas linked to driverless vehicles. In this way, through the design of hypothetical scenarios and the analysis of the platform participants' average behaviour, the *Moral Machine* algorithms can better understand the context, underlying conflict and expected *behavioural responses* when confronted by a similar situation as part of the platform *training programme* or even in a real scenario. This type of combined human-machine exercise is called *extended intelligence* (Ito 2017, 2019) and, in this particular case it is linked to critical studies on the rationalization of machines' moral behaviour: *extended ethics*.

> Instead of thinking about machine intelligence in terms of humans vs machines, we should consider the system that integrates humans and machines—not Artificial intelligence but extended intelligence. Instead of trying to control or design or even understand systems, it is more important to design systems that participate as responsible, aware and robust elements of even more complex systems (Ito 2017, 2019).

Given that Artificial intelligence plays an increasingly important role in *autonomous driving* technology, it is believed that experiments such as *Moral Machine* (Awad et al. 2018) can help find plausible solutions to the ethical conflicts underlying driving, and which driverless vehicles will have to face, by using democratic platforms intended for content design, data generation and training, as well as the progressive application of knowledge.

Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations, it will happen in that most mundane aspect of our lives: everyday transport. Before we allow our cars to make ethical decisions, we need to have a global conversation to express our preferences to the companies that will design moral algorithms and to the policymakers who will regulate them.

Moral Machine was rolled out to begin such a conversation, and millions of people weighed in from around the world. Respondents could be as parsimonious or thorough as they wished within the ethical framework they decided to follow. They could engage in a complicated weighting of all nine variables used in *Moral Machine*, or adopt simple rules such as "let the car always go onward". Our data helped us to identify three strong preferences that can serve as building blocks for discussions of universal machine ethics, even if they are not ultimately endorsed by policymakers: the preference for sparing human lives, the preference for sparing more lives, and the preference for sparing young lives. Some preferences based on gender or social status vary considerably across countries, and appear to reflect underlying societal-level preferences for egalitarianism (Awad et al. 2018: 63).

The interest in moral machine learning extends beyond the driverless car sector. In the sphere of democracy, the prominent role given to Artificial intelligence in the decision-making processes of governments, institutions, organizations and related companies is enabling a greater optimization of resources; market behaviour forecasts; our understanding of the expectations and interests at stake on a particular issue; adaptations to legal, social and moral changes; approaches to managing corruption, nepotism and disaffection; and methods of attracting talent, knowledge and funding, among other things. As a result there has been an increase in specific proposals for algorithmic politicians to occupy high positions in local and national governments and likewise for algorithmic CEOs, managers and middle managers to manage and shape the direction of media and financial organizations.

As was seen in Chap. 3, in the political sphere the deployment of artificially intelligent mathematical models in public policymaking and management is a growing phenomenon. In this respect, key examples are provided by the *political algorithms* discussed above, AI Mayor —the algorithm that came third in the mayoral elections in Tama City (Tokyo) in April 2018—SAM (Semantic Analysis Machine) and Alisa —the two algorithms that started their political careers in 2017 with the intention of running in the New Zealand 2020 and Russian 2018 presidential elections respectively—, as well as Leader Lars —the algorithm that ran for president of Denmark in 2022—and the algorithm that was officially appointed advisor to the president of Roman in 2023.

Similarly, there has also been a significant use of algorithms in democratic governments. For example, as was revealed by the New Zealand government's *Algorithm Assessment Report* (Stats NZ 2018), algorithms play an essential role in the services the government delivers to citizens and they also contribute to the development of new, innovative and targeted policies that are intended to

enhance the government's capacity to meet its goals.

Deep Knowledge, whose contribution in the economic sphere was discussed in Chap. 2, is a valuable example. In May 2014 this investment fund working in the biotech sector appointed an artificially intelligent algorithm as chairman of its board of directors. VITAL (Validation Tool for Advanced Life Sciences), as this digital CEO is called, performs advanced data analysis of scientific publications, grants, patent applications and clinical trials in order to decide, along with the other board members, on issues such as where the company should invest. However, in this case, the algorithm's opinion carries greater weight, as it has a casting vote in the event of a tie (Burridge 2017; Pardo 2014). Other significant cases are companies such as Xerox, Google, Unilever, L'Oreal and Amazon. All these companies have assigned an artificially intelligent algorithm to work in or serve as the head of their human resources departments (Walker 2012). Their role is to improve recruitment, planning and personnel management processes and thereby find ideal candidates, forecast, and locate conflicts, while also preventing any drain of talent (Sagristà 2016). Xerox uses and markets a workflow automation tool for recruitment and selection based on an artificially intelligent algorithm (Xerox 2018); Amazon acknowledged in 2018 that it has been using an artificially intelligent algorithm for recruitment and dismissal since 2005 (Dastin 2018); L'Oreal uses an artificially intelligent algorithm for its 'The Beauty Lab' scholarship programme to select the best profiles for marketing, trade marketing, communication and digital (L'Oreal 2018); Unilever has commissioned an artificially intelligent algorithm to recruit and hire future leaders as part of its *Future Leader's Programme (UFLP)*; and, finally, Google uses an artificially intelligent algorithm to predict which candidates are most likely to achieve business success.

And, in the media sector, the Los Angeles Times, Forbes, the BBC and Xinhua all stand out for this practice. All of these different media organizations are implementing intelligent algorithms to interpret and convert data and metadata into information, to manage information flows and corporate communication, to write news stories and produce chryons, as well as presenting television news programmes. The Los Angeles Times, the fourth most widely read newspaper in the US, has been using an artificially intelligent algorithm called Quakebot to write news stories since 2014 (BBC 2014); Forbes, a magazine specializing in business and finance, and the BBC, the UK's public service broadcaster, have been using an artificially intelligent algorithm called Quill to write news articles and chyrons for news programmes for years (Marr 2016); and Xinhua, China's state news agency, has been using two *AI anchors* named after their real-life counterparts Zhang Zhao and Qiu Han on their

programme *China Xinhua News* since 11 November 2018 (Vidal 2018).

Nevertheless, despite the benefits of their application and use in a range of contexts, these artificially intelligent mathematical models are the subject of harsh criticism and have prompted major doubts, both from within their organizations and beyond. All too often decisions made by these mathematical models have revealed a clear bias towards men, rich people, heterosexuals and Caucasians to the detriment of women, poor people, homosexuals and African-Americans, to name just a few cases. In this way, they encourage exclusion, widen the inequality gap and reproduce models of behaviour that society has been trying to eradicate for decades (Calvo 2018; O'Neil 2016). As studies have shown, decisions based on artificially intelligent algorithms: tend to favour the rich and punish the poor during the design and implementation of public policies (O'Neil 2016); exclude women from recruitment processes (Rubio 2018); favour the social integration of white criminals over black criminals (ABC 2018); link women to housework and men to liberal professions; confuse black people with gorillas, and so on. Evgeny Morozov has harshly criticized their use and stated that "(...) information-rich democracies have reached a point where they want to try to solve public problems without having to explain or justify themselves to citizens. Instead, they can simply appeal to our self-interest—and they know enough about us to engineer a perfect, highly personalized and irresistible nudge" (Morozov 2013).

The increase in cases of malpractice resulting from a lack of objectivity, high degree of opacity, behavioural relativism and biased arbitration, among other things, reveals levels of technological immaturity that underscore the need to rethink the excessive expectations generated by some intellectuals, professionals and specific groups regarding the application of Artificial intelligence in the short term in order to avoid the negative consequences that it is currently producing or may go on to produce. It is for this reason, as is the case for driverless cars, that one of the key issues being explored in the design and implementation of decision-making processes is how to make it possible for artificially intelligent machines to learn to make morally valid decisions.

Work on this issue has gone in two directions. On one hand, it has explored the ethics of Artificial intelligence with the main aim of providing AI with a normative framework for the design, application and just and responsible use of mathematical models, while also providing the STEM professionals who work in this field with an ethical code to work with. On the other hand, work has been undertaken on algorithmic ethics, seeking to endow artificially intelligent mathematical models with the moral—communicative, affective, cognitive, hermeneutic and ontogenetic—competences and capacities needed to enable

them to make just and responsible decisions. With regard to the former, work has been undertaken on the ethics of Artificial intelligence, and a number of interesting governmental, institutional, academic and civil society proposals have also been made. The latter include the creation of guidelines, codes and ethics committees, and the study of new areas of applied ethics (digital ethics, roboethics, ethics of robotic intelligence, etc.) and social responsibility (technological social responsibility, robotic social responsibility, social responsibility of Artificial intelligence and so on). With regard to the latter area, research has been undertaken on artificial neural networks and Deep Learning in order to explore algorithmic ethics, addressing issues such as the development of ethical patches for moral learning.

6.2 From Artificial Neural Networks to Artificial Moral Intelligence

As a field of research, AI is currently at its highest peak since its emergence in the summer of 1956. It was then that ten researchers interested in neural networks took part in a six-week workshop in order to find ways of enabling machines capable of using language and forming abstractions and concepts to solve problems hitherto reserved for humans, while also improving their own operational capacity (Bostrom 2014).

As has been discussed above, during the 1980s AI was on the verge of succumbing to the lack of interest of its target audience in the wake of forecasts that its advances and the expectations of potentially major practical applications were in fact unrealistic (Bostrom 2014). The strong comeback AI has made today is, above all, due to the emergence of neuroscience. Study in the latter field has led to a deeper understanding of neural learning processes, and in turn the consolidation and expansion of the Fourth Industrial Revolution. Its influence on sectors such as retail, cosmetics, communication, education and healthcare has revealed unprecedented potential for optimizing scarce resources and improving manufacturing, communication, education, healthcare and decision-making processes involved.

Neuroscience is a branch of physiology that studies the cerebral foundations of human or animal behaviour by using non-invasive techniques to capture images of its activity, for example Functional Magnetic Resonance Imaging (fMRI), Repetitive Transcranial Magnetic Stimulation (rTMS), Magnetoencephalography (MEG) and Positron Emission Tomography (PET). The main aim of this field of study, which emerged at the end of the 1990s, is to

identify the cerebral foundations for human behaviour, especially those involved in decision-making processes used in a range of contexts of activity. The increased knowledge of the brain and the enormous amount of data produced by neuroscientific studies has led to major advances in AI. On one hand, these studies are fostering advances in the design of artificial neural networks, and on the other, they have become an important source for amassing a high volume of data on emotions, feelings, values and other factors and elements involved in intelligent machines' learning and decision-making processes.

However, with regard to the application and use of AI in the political and economic sphere, as has been discussed, it is algorithmic governance that attracts the most attention. The expectations currently being discussed are so optimistic —the eradication of corruption, nepotism and selfishness in decision-making processes, resource optimization, increased competitiveness, enhanced profits, and so on—that as a result business managers and/or political representatives are being replaced by mathematical decision-making models. Nevertheless, it is one thing to be able to process massive volumes of data concerning the preferences, opinions and habits of the people involved in any one activity for the purpose of generating statistics on a range of issues, but it is quite another to actually govern those areas of life. Among other reasons, this is because governance requires leadership, autonomy, prudence, justice, a sense of the common good, recognition, tolerance and accountability. However, decision-making algorithms have not been able to demonstrate their grasp of the minimum competences and capacities for learning and moral development, and it is clear that *Artificial Moral Intelligence* (AMI) is currently facing a range of problems:

- Firstly, with regard to morality as a competence, or in other words, the possibility of developing a structure capable of making judgements about value conflicts, learning from experience and making moral decisions.
- Secondly, concerning morality as a capability, which may be summed up as a machine's knowledge of how, when and why it should use its artificial moral structure for resolving value conflicts or else simply making decisions.
- Thirdly, the issue of morality as learning, which refers to whether it is actually possible that machines can elaborate, learn, assimilate and develop critical and post-conventional moral knowledge.

These three issues alone are quite enough of a handicap for Artificial intelligence. Moral conflicts are problematic and demand moral plurality, discursive processes, universalizable values, post-conventional viewpoints, a constant critique, individualized case analysis and a choice of possible courses of action. By contrast, current moral machine learning (MML) is dilemmatic,

descriptive, comprehensive, inconsistent, empirical, decisionist, positivist, vicarious, naïve and immoral. As Diego Gracia has stated, approaching moral conflict in a dilemmatic manner gives rise to unrealistic situations and on the whole immoral responses (Gracia 2001); hence, efforts are currently underway to develop the moral competences and capabilities of AI-enabled machines.

A fundamental contribution to this latter development is being made within the field of neuroscience. Among other things, neuroscientific studies have become a key element in the development of Artificial intelligence (Nguyen 2017). Thanks to the range of data and information contributed by neuroscientists on how the brain functions in a range of human contexts, and also during decision-making processes, improvements have been made to the internal design of mathematical models. This has allowed the enhancement and optimization of their predictive and decision-making capabilities.

Firstly, the increased knowledge about human behaviour contributed by recent neuroscientific advances human has led to an unprecedented development of Artificial intelligence. Currently, what is known as *machine learning*—algorithms based on a set of given rules that allow them to make autonomous decisions in order to provide responses modified to address practical problems—has evolved into *Deep Learning* (Bostrom 2014; Iba 2017): an algorithm-based structure formed of artificial neural networks. The latter are computational models, such as Google's *Neuronal Machine Translation* (NMT), which mimics the operation of the nervous system of a living organism. These ANNs do not make decisions based on a watertight model containing a repertoire of pre-established strategies and responses. Instead, it is the algorithm that learns from experience—continuous trial-and-error processes, for example—in order to establish and/or reinforce connections between parameters or between different layers of processing units (artificial neurons that specialize in the analysis of a particular feature of the observed object). This enables it to improve its abstractions and perceptions of reality and thus enhance the results it is expected to deliver (Conde et al. 2018; Hirasawa et al. 2018; Webb 2018; Higham and Higham 2018).

Furthermore, neuroscience has become an important and inexhaustible source of data for improving the predictability of *connected things* and streamlining the decision-making processes used in algorithmic models. For example, studies undertaken in the different neuroscientific subdisciplines, such as neuroeconomics, neuroethics, neuromarketing or neuropolitics, are continuously providing huge amounts of *objective* and *measurable* data on human behaviour derived directly from analysis of the brain, and can be processed by *Big Data* analytical tools to improve the predictions and decisions

made by mathematical models and therefore minimize the underlying uncertainty of *cyber-physical systems* (Chen and Lin 2014; Higham and Higham 2018; Marblestone et al. 2016).

Undoubtedly, the alliance forged between neuroscience and the technological convergence of *Big Data*, IoT and AI is changing the way things are done. However, the consequences of these changes are not always good or acceptable to a large part of those affected by the system (O'Neil 2016). Political and economic neuromarketing, for example, can use *cyber-physical ecosystems* to design and implement highly addictive campaigns capable of modulating and/or manipulating the free will of the system's subjects. Similarly, in order to maximize a particular benefit obtained by a small minority, they are capable of diminishing and/or inhibiting the capacity for critical judgement exercised by voters, rulers and customers (Conill 2012; Conill and García-Marzá 2012). The same systems also allow political and economic institutions and organizations to monitor even the private and intimate behaviour of voters, customers and competitors in order to predict and control their actions and decisions, as well as creating attractive or addictive products and implementing decision-making processes based on such data.

The solution to this problem appears to be to develop sufficient technological competence, capacity and moral learning to enable machines to make just and responsible decisions. A variety of models of moral machine learning have been proposed and developed in recent years, and the next section focuses on a survey of these.

6.3 Moral Machine Learning Models

One of the main concerns facing AI today is how machines can make morally valid decisions. The development and deployment of driverless cars, for example, has been slowed and even halted by the difficulties faced by designers of artificially intelligent algorithms when it comes to deciding and integrating the most appropriate moral rationale criteria to tackle or overcome the value conflicts underlying smart mobility in practice. Among other issues, a key question is when to put the lives of vehicle passengers at risk in favour of pedestrians, or vice versa, in the event of an accident. The variety of applications of algorithmic governance also generates a great deal of mistrust due to the continuous evidence of malpractice deriving from the emotional, misogynistic, xenophobic, homophobic and aporophobic bias that is inherent in the current artificially intelligent algorithms. There is also the issue of their lack of accountability, as discussed in Chaps. 4 and 5.

To address such challenges, Artificial intelligence has focused on three main lines of research for the purpose of developing *deep neural networks* capable of interpreting diverse contexts of reality, grasping the meaning of morality and acting accordingly: learning and internalizing values; designing and introducing *ethical patches* into mathematical codes; and acquiring hermeneutic, argumentative and emotional competences and capacities.

As Bostrom in particular has discussed regarding the first line of research (2014), there are currently five relevant proposals for how AI can undertake a process of acquiring and internalizing moral values, and this provides a frame of reference for machines' decision-making: evolutionary selection; reinforcement learning; learning by association; motivational scaffolding; and evaluation.

(a) *Learning through evolutionary selection*: this aims to emulate the process of biological evolution for acquiring values by using search algorithms which, on one hand, act by expanding the catalogue of values and, on the other, undertake the selection and enhancement of the best values while also identifying and eliminating the worst: in other words, those that function best and worst in the real world. This two-way process can be equated to what the biologist John Maynard Smith (1982) termed an *evolutionarily stable strategy* (ESS) and *replicator dynamics* (RD), except that it addresses values rather than social behavioural strategies. Essentially this is a search algorithm, which, on one hand, collects the values from the real world that are used to resolve moral conflicts and remain unchanged over time (expansion); it then establishes an order of materiality for the purpose of eliminating the values that score lowest in a test performed as an evaluation function (contraction). Finally, it processes the available information in order to mutate or specify the optimized final values for achieving the intended aims. The problem raised by this kind of value assimilation is that, as Bostrom argues, "Nature might be a great experimentalist, but one who would never pass muster with an ethics review board—contravening the Helsinki Declaration and every norm of moral decency, left, right, and center." (Bostrom 2014, 188). Moreover, it must be borne in mind that this is a conventional and strategic proposal that instrumentalizes values, and the concept of morality adhered to is based on a functionalist criterion. The criterion of choice and the acquisition of values are not linked to an enhanced orientation of action in any specific case, but rather to becoming better adapted to the environment. Furthermore, the lack of any critique during the whole process generates a

purely conventional axiological framework which can give rise to and perpetuate injustices.

(b) *Reinforcement learning:* this involves the algorithm receiving some kind of reward, assessment, encouragement or information (from the programmer or the environment, for example) about the decisions it has made. To achieve this, an evaluation or utility function is assigned which, through experience (trial and error), allows the algorithm to refine and improve its estimates of the practical use of certain values, associate them with responses and behaviours that maximize benefit, and, finally, reinforce or correct its decisions. For example, a reward (one point) is given for each decision based on certain values, but the algorithm is not informed why it is being rewarded, nor what values it is being rewarded for using. However, the main problem with this type of reinforcement learning is that, as Samuel Bowles argues, good incentives do not make good citizens (2016). As has been shown in field studies and laboratory experiments performed with strategy games, over time the beneficiaries of incentives become unable to act morally unless there is a reasonable expectation that they will be rewarded for it. Furthermore, behaving morally does not always produce positive intrinsic or extrinsic rewards for individuals. Not infrequently, doing the right thing comes at a high cost, such as reduced adaptation, reduced personal well-being or reduced benefits. The problem is that human beings not only have a moral structure, they also have the capacity to critique what is conventional in order to discern what is just through dialogue and by forming agreements with all concerned.

(c) *Value accretion:* this is concerned with associating values with concepts, things, processes, experiences or objectives, just as happens when, for example, human beings confer a set of values on concepts such as a person, profession, institution, etc. With regard to Artificial intelligence, the aim is to select, link and adapt complex values to given aims, i.e. governing with leadership, respect, honesty and responsibility. However, the challenge that arises is the difficulty involved in forming an in-depth understanding of the mechanism that enables humans to associate values with concepts, as well as the fact that a neurocognitive architecture can only be applied to a complete brain emulation. By contrast current Artificial intelligence is only capable of a very small emulation of the human brain (Bostrom 2014). Moreover, firstly, it is not clear what reference would be used by an Artificial intelligence system to determine the values that should be

associated with a given concept, as well as critiquing them and, if necessary, changing them for others. Secondly, it is also unclear what motive an AI would have for choosing, associating and being guided by these values without any reinforcement of its utility function. On the basis of all the studies conducted to date it would seem that, as with the previous proposals, the reference used by this form of AI is once again functionality —the values best suited to attaining a given aim—while the motivation is profit maximization—whatever gives rise to the achievement of the given aim. Thus, learning by association once again proposes an instrumentalization of moral values accompanied by moral foundation based on functionality.

(d)

Motivational scaffolding: this seeks to establish methods for selecting goals, aims, motivations and guidelines. On one hand, this would involve equipping the AI with an instructional scaffolding consisting of simple provisional aims and final instrumental aims that would serve as a horizon of meaning that could be used to pursue instrumental aims. However, this must be simple enough to grant the AI the possibility of developing them and, if it deems it necessary, modifying them. On the other hand, it involves providing motivational scaffolding for pursuing interim aims and objectives. Therefore, it seeks to provide the AI with a learning structure consisting of a minimum content which, on the basis of experience, the AI modifies as it becomes capable of dealing with scenarios of increasing complexity. However, while instructional scaffolding is a highly interesting area of study, its implementation is currently hindered by its lack of a stable foundation. Its limitations include the possibility of the programmer having too much control and influence, which runs contrary to current ideas of Artificial intelligence. When implementing instructional scaffolding, it is the programmer who, to a large extent, creates the primary horizon, albeit a simple and open one, which frames the development and modification undergone by the AI in response to experience acquired through the pursuit of its interim goals. This likewise defines the motivations the AI responds to.

(e)

Value learning: this approach seeks to provide the AI with a criterion for independently learning the values it is expected to use and define based on what it calculates to be the best way to act according to these values. Various difficulties are raised by this proposal. Firstly, how should decisions be made about which criterion should be used? There is also the

> fact a *strong AI* is needed: one that has evolved sufficiently to enable it to detect and understand an environment's axiological structure by making increasingly precise hypotheses, which are in turn based on estimates founded on the AI's principles along with all available empirical data. Furthermore, as with all the other proposals, it remains unclear how it is possible to ensure "(...) that the AI would be motivated to pursue the programmed values in the way we intended" (Bostrom 2014, 197); or in other words, the AI may know and be able to perfectly interpret its axiological framework, yet that does not mean that it will adhere to it. Finally, it should be borne in mind that the AI's estimates of values, as well as its intermediate and final aims are based on what is, not what should be.

A common feature of all these different approaches to the acquisition and learning of values by AI applications is that they are strictly conventional and framed by strategic interests. They are based on the uncritical observation of the patterns, opinions and behaviour belonging to a particular community or else a programmer's view of what is fair and good for society. However, a major problem is that moral values cannot be based on conventions or strategic interests. Moral values are post-conventional, and hence, as Adela Cortina argues, they are rooted in the terrain of intersubjectivity (2010).

With respect to the second of these five lines of research, *Reinforcement learning* —the design and introduction of *specific codes* or *ethical patches* into mathematical code—the studies such that have been undertaken, in particular by Bertram F. Malle (2016), have identified the multiple elements that make up human moral competence: a moral vocabulary; systems of norms; moral cognition and effect; moral decision-making and action; and moral communication. They propose, on the whole unsuccessfully and with little specific detail, how to design robots that possess one or more of these human aspects. Michael Kearns and Aaron Roth (2020) have contributed another noteworthy proposal, the *ethical algorithm*. They argue that it is feasible to extend the principles machine learning is based upon to include the moral values that prevail in societies, because "(…) all decision-making —including that carried out by human beings—is ultimately algorithmic. The difference is that human decision-making is based on logic or behaviors that we struggle to precisely enunciate" (Kearns and Roth 2020, 190). Thus, they believe, if we were able to describe this process with sufficient accuracy, we would represent it by means of computer algorithms. It is not a question of whether algorithms should be avoided, but of how to manage to use precisely defined algorithms (Kearns and Roth 2020, 190–191). However, their proposal and discussion does

not go beyond setting out ideas on anonymizing data and preserving privacy.

Finally, with regard to the third line of research, *Learning by association*, which explores the acquisition of hermeneutic, argumentative and emotional competences and skills, the results achieved are not overly significant and are linked to *weak AI*. However, the advances made with regard to dialectical capabilities are worth noting. These include IBM's Project Debater: the algorithm that in 2018 won a debating competition against a person for the first time, and in 2019 came close to beating the world debating champion, Harish Natarajan (Expansion 2019). A *learning by association* study has also been conducted by the Centre for Cognitive Science at the Technische Universität Darmstadt (Jentzsch et al. 2019), having incorporated a large corpus of written text into a *deep neural network* in order to instruct artificially intelligent machines to learn vector representations of words, has demonstrated that these AI systems are capable of extracting arguments concerning correct and incorrect behaviour from texts. Finally, advances have been made concerning the capacity of intelligent machines to recognize emotions in humans, mainly through physiological "empathic" reactions (Boden 2017, 76–77), and there have also been significant studies devoted to *healthcare robots* equipped with Deep Learning. Their tasks require them to engage in social behaviour—that is, to interact and collaborate proactively with humans—as part of correctly performing their role.

However, as Bostrom (2014, 207) states, "Goal system engineering is not yet an established discipline. It is not currently known how to transfer human values to a digital computer, even given human level machine intelligence." Worse still, even if in the future it is discovered how Artificial intelligence can collect and store values, there are many issues related to human morality that are not even being addressed. These include the reciprocal requirement to recognize the communicative and affective capacities of the parties engaging with each other; the ability to discern between applicability and validity, as well as common sense and the possibility of reproducing and developing a conscience, motivation, voluntariness, emotionality and moral responsibility. Essentially, it is one thing to have the competence to acquire a set of prevailing social values, but it is quite another to become competent and capable of criticizing them in order to validate, revise or reject them; to be moved by their practical occurrence; to have the sense of a bad conscience or indignation when they are violated; to appreciate their occurrence and monitor them, and demand appropriate behaviour from oneself; or to be accountable for the decisions we make and the behaviour we engage in, among many other things.

6.4 The *Non-place* of the Moral Machine

As Morozov has warned, the dominant tendency of our digitally hyperconnected society is to recreate a heteronomous society that places political and economic, as well as everyday decision-making processes in the hands of algorithms.

> Thanks to smartphones or Google Glass, we can now be pinged whenever we are about to do something stupid, unhealthy, or unsound. We wouldn't necessarily need to know why the action would be wrong: the system's algorithms do the moral calculus on their own (Morozov 2013).

This new reality, based upon the ongoing digital transformation, increases people's vulnerability and, therefore, proportionally increases the demand for moral knowledge, whose social function is, precisely, to minimize vulnerability and its consequences by providing resources to guide praxis and thereby avoid the emergence of conflicts, while also coordinating actions intended to attain common aims and resolve the conflicts that underlie political and business practice (García-Marzá 2016, 882–883; 2005, 254–256). In other words, the community of experts working on machine learning and moral behaviour do not show any significant interest in the moral knowledge the system's participants intuitively use of during a dialogue intended to achieve an intersubjective understanding about something in the world. Instead this now seems to be reduced to socially atomized courses of action and mathematical models capable of collecting and processing the quantifiable and available information. There has therefore recently been an exponential increase in the attention devoted to Artificial intelligence by scholars working in the field of ethics. Three concerns —structure, content and moral motivation—stand out from the various critiques of the possibility of designing AI-equipped machines being capable of learning to make moral decisions in the sphere of democracy.

Firstly, current proposals for moral machine learning are above all concerned with the application, development and implementation of a mechanism to allow the acquisition of values, but they pay very little attention to other important and necessary parts of the moral structure of human beings. Little or nothing is said, for example, about how to create an artificial moral structure capable of, among other things: emulating moral judgements about what is just or unjust; distinguishing between moral and social values; interpreting and critiquing both knowledge and action; generating *prosocial emotions* capable of promoting adherence to values independent of the benefits derived from their

implementation; or possessing common sense when applying moral norms and values to concrete contexts. Many of the current proposals therefore create a kind of substitute for artificial moral judgements based on a goals-focused utility function, rather than adhering to principles, values or norms whose consequences could be accepted by all those affected. This would in turn be framed by an overarching practical discourse with specific logical rules and a procedural moral principle (Cortina 1993: 208).

Secondly, the current proposals for moral machine learning demonstrate little concern about the content of morality, above all with regard to how to specify, justify and enhance the values, norms, principles and moral virtues that enable the structure to function correctly and legitimately. This disinterest appears to be the result of a lack of trust in the current state of AI, which explains the excessive paternalism encountered in the majority of the proposals. It is striking that, while extolling AlphaGo's autonomous and creative capacity to design and implement strategies that have never been thought out before by humans, the potential creative capacity of this AI is controlled with patches intended to prevent it from enriching existing values with arguments never before proposed by humans. It seems clear that AI experts acknowledge the impossibility of addressing the complexity of morality through mathematical formalism, as well as the incompatibility between computational rationality and practical rationality. Furthermore, there are clearly good reasons to be deeply sceptical of claims that it will become possible to leave the course of humanity in the hands of AI. So, in response to the widespread distrust of AI, programmers have had to insert *patches* to prevent a possible *brain puncture*, in addition to including *utility functions* based on aggregates of scores achieved by pursuing intermediate goals and profit maximization in order to avoid the possibility of goals being abandoned. This enables AI experts to implement a quasi-scrupulous monitoring of the assigned aims and standards.

Finally, in human beings the motivation behind wanting to acquire and uphold moral principles, values and norms emanates from the sense of duty to do what we have good reason to consider just and desirable. However, Artificial intelligence builds a motivational structure based on the degree to which moral values and norms are involved in achieving the given goals either through a pre-specified utility function, or the appraisal of moral values undertaken by programmers. The criterion on which decisions are based for the majority of value acquisition and learning techniques is therefore profit maximization, which means that AI instrumentalizes moral values to improve its results. AMI therefore has a major deficit in terms of emotional motivation, which makes its design and development difficult or even impossible.

In short, as Searle warns, it is worth bearing in mind that there is always a temptation to think that the solution to our problems depends upon an until now uncreated technological marvel. However, the nature of refutation is completely independent of the current state of technological development. It has nothing to do with the precision of a digital computer, nor what a digital computer is (Searle 1988: 36).

Moreover, as Margaret A. Boden has argued, “To think that you can solve human conflicts such as the Middle East (or Northern Ireland or Catalonia) with AI is utterly ridiculous. Those who believe in [*technological*] *singularity* [the stage at which the development of machines’ intelligence surpasses that of humans and becomes uncontrollable and irreversible] ignore the limitations of current AI. Their beliefs rely solely on exponential technological advances, but ignore a key fact: increasing computing power and data availability does not guarantee the attainment of an AI equivalent to human intelligence” (Fresneda and Alba 2018).

Moreover, beyond the issue of *technological singularity*, consideration must be given to the consequences of applying and developing *weak AI*. Pierpaolo Donati, for example, has argued that this is giving rise to a socially divided and highly conflictive ecosystem: “While some celebrate the new world of digital technologies (ICT, Artificial intelligence, robots) as the arrival of ‘augmented man’, others think that this world is leading us towards further dehumanization, in the sense that, in contrast to the human, it results in a social life ‘without quality’” (Donati 2019, 31–32).

Therefore, looking beyond attempts to generate an algorithmic moral intelligence, governmental bodies, institutions, public and private organizations and companies, as well as civic groups, are endeavouring to design and develop the practical application of a form of ethical governance with regard to AI, which would enable the design, application and use of artificially intelligent mathematical models in a just and pleasing way.

6.5 Towards the Ethical Governance of Machines

The new socio-political, economic and cultural contexts that have grown up around Artificial intelligence, along with the digital techniques and technologies it seeks to develop into practical applications, offer major opportunities for social development across a range of fields of activity. However, they also make people increasingly vulnerable due to the immense complexity of digital contexts and their negative and, in many cases, uncontrolled impact. This is the paradox underlying the current digital transformation processes, and addressing

it requires an increased and wider use of moral knowledge. The social function of this must be, above all, to minimize people's vulnerability by providing resources to guide technological practice and prevent the emergence of injustices in the diverse contexts this technology is used in.

In response to the challenge Artificial intelligence poses for mature democracies, as well as the different techniques and digital technologies involved in decision-making processes, along with their current or virtual impact on society, over the last five years various governmental bodies, public and private organizations, associations and foundations of different levels have drawn up and implemented laws, decrees, standards, guidelines and codes at corporate, national and international level. In all of these, special emphasis is placed on ethical and legal factors in order to develop solutions to the current virtual problems raised by the impact of the development and application of AI across a range of fields of activity, such as economics, politics, health or communication.

In this regard, as has been shown in the study "Ethically governing Artificial intelligence in the field of scientific research and innovation" (González-Esteban and Calvo 2022), the most noteworthy developments have been legislative initiatives–such as laws, decrees and regulations; self-regulation–guidelines and codes of ethics, conduct, professional deontology and good practices and accountability reports; and ethical governance systems, which have been developed around the world. Nevertheless, two key issues stand out (Jobin et al. 2019). Firstly, almost half of these documents were drafted by private companies (22.6%) and government agencies (21.4%), while the rest were produced by research institutions (10.7%), intergovernmental or supranational organizations (9.5%), NGOs (8.3%); professional/scientific associations (8.3%), private sector alliances (4.8%), research alliances (1.2%), scientific foundations (1.2%), trade union federations (1.2%) and political parties (1.2%). Secondly, the analysis and traceability of the different ethical principles included in these documents can be reduced to 11 fundamental principles for AI: justice and welfare; non-maleficence; responsibility; privacy; beneficence; freedom and autonomy; trust; sustainability; dignity, solidarity and explicability. Essentially, this comprises 10 traditional principles which are widely used in the range of contexts, and one original one: explicability.

With regard to legislative initiatives, laws, decrees and regulations are being designed and implemented by government agencies to control the design, application and impact of Artificial intelligence. In the European Union, for example, two proposals stand out. Firstly, there is the *Civil Law Rules on Robotics* (European Parliament 2017). This is essentially a regulation that

revises and extends the four *Laws of Robotics* set out and developed by Isaac Asimov over 50 years ago (Asimov 1942, 1950, 1982, 1993) to include current concerns such as transparency, confidentiality and accountability in the design, application and use of intelligent robotics. On the other hand, the *Artificial Intelligence Act* (European Parliament 2021) marks another significant development. The main aim of this regulation is to ensure the safe and socially acceptable development and application of Artificial intelligence.

Nevertheless, while these regulatory initiatives are a significant development, their impact is expected to be limited for a number of reasons. Firstly, this is due to the speed at which Artificial intelligence and the techniques and technologies for its practical application are currently being developed. As a result, any legal-political framework must undergo constant, hasty revision to adjust it to new developments and prevent its design from becoming anachronistic, as well as ensuring its capacity to influence is not left without effect. Secondly, the legislative process itself is a limitation. The nature of this process makes it very difficult to bridge the gap between detecting changes and their impact and the subsequent revision of the legislative framework in order to control them. A final hindrance is the nature of the state that characterizes the legislative frameworks. State factors limit the legislation's scope and limits its results, because processes of digital transformation are characterized by hyper-globalization, which requires supra-state responses.

With regard to self-regulatory initiatives, the most relevant proposals of those developed to address the development and implementation of guidelines and codes of ethics, conduct, professional ethics and best practices for the purpose of guiding the design, application and use of Artificial intelligence are: the establishment of ethics committees for resolving AI-related conflicts and the publication of accountability reports on the impact of the application of Artificial intelligence upon the market or civil society. In this case the American continent has gone further, and in the USA alone over 20 proposals, from different organizations, associations and, above all, companies, have been made for guidelines and codes of ethics, conduct or good practices for the design and use of Artificial intelligence (Jobin et al. 2019). The same is true for the European continent, where the *High-level expert Group on Artificial Intelligence*, whose participants are mostly representatives of large private corporations, has developed the *Ethics Guidelines for Trustworthy AI* (2019). The aim of these guidelines is to establish a set of core principles for Artificial intelligence that can serve as a guide for the elaboration and implementation of laws, codes, conduct and good practices.

Nevertheless, self-regulatory initiatives such as this have prompted

scepticism and criticism due to the degree of self-interest motivating them (Cotino 2017). It seems that it is concerns about the moral orientation of the design, application and use of Artificial intelligence do not lead many organizations and companies to develop and implement such initiatives. Instead, they present arguments against any unfavourable state regulation which would limit the major profits they obtain or can obtain from AI. One clear example of this was the failed launch of Google's *Advanced Technology External Advisory Council* (ATEAC). This AI ethics committee was disbanded after just 2 days of having been formed in the wake of a surge of unrest among Google employees and the wave of resignations by committee members after it was revealed that their fellow members included avowedly denialist, homophobic, xenophobic and misogynist individuals (Bietti 2020), such as Kay Coles James of The Heritage Foundation, as well as representatives from the arms industry, such as Dyan Gibbens of Trumbull, who upheld adamantly immoral and warmongering interests.

Finally, with regard to ethical governance systems, a number of interesting but insufficient proposals have been made which, to a greater or lesser extent, attempt to engage with and resolve the problems underlying the development, application and use of Artificial intelligence in fields of activity. A number of initiatives promoted by the European Union stand out. Firstly, a governance proposal has been made in the report entitled *AI Watch—Artificial Intelligence in public services* (Misuraca and van Noordt 2020). However, this is focused upon the field of public administration and it explores a dilemma perspective and an ethically insufficient concept of "governance 'with and of' AI". Secondly, Title VI of the aforementioned *Artificial Intelligence Act* (European Parliament 2021), establishes a clearly insufficient governance system which is merely based on the establishment of a European Artificial Intelligence Committee for the purpose of ensuring that the application and implementation of the Regulation is complied with, in addition to encouraging the exchange of good practices. The same may be said about the *Data Governance Act* (European Parliament 2020). Although, it proposes introducing certain technical instruments to ensure the protection, privacy and confidentiality of data when transferred, reused and evaluated by third parties, its proposal for horizontal governance solely addresses compliance with the current legislation, and it deliberately excludes those affected from the entire participatory process. Finally, the proposal entitled *Responsibility Research and Innovation* (RRI) (European Commission 2012), which focuses solely on the field of research and innovation, is founded on a problematic view of moral conflict and a concept of "Science with and for society" that is deeply concerned with the issues of the inclusion and active

participation of those affected in the assessment, development and decision-making processes linked to scientific research (European Commission 2012).

Despite the volume and interest demonstrated by the range of initiatives emerging from the state, the market and civil society, there can be no doubt that there is still a great deal of work to do to establish efficient control over and provide proper guidance for the processes of designing, implementing and using Artificial intelligence across a range of areas, such as democracy. Being able to place trust in Artificial intelligence is a necessary and unavoidable stage towards achieving the aims and expectations identified with it across all these spheres; undoubtedly, legislative frameworks, self-regulatory proposals and governance systems are plausible tools to generate and develop it. Nevertheless, a necessary preliminary step is to develop a deeper understanding of the core problems that undermine such initiatives. Among other things, it must be asked what role should be played by those affected by the consequences derived from the design, application and use of Artificial intelligence? Who should decide the principles and values that constitute the axiological framework that enables the rules, actions, decisions and impacts linked to Artificial intelligence to be prescribed and critiqued? And how should this be decided? Which communication tools enable effective reliability management? And finally, how can all these issues be systemized, applied and implemented in areas as complex as the democratic sphere?

Bibliography

ABC. 2018. Cuestionada la máquina que predice futuros delitos en EE.UU. *ABC*, 01 January 2018. https://www.abc.es/ciencia/abci-cuestionada-maquina-predice-futuros-delitos-eeuu-201801181039_noticia.html.

Asimov, Isaac. 1942. Runaround. In *Astounding Science*, ed. J.W. Campbell, 94–103. New York: Fiction, Street and Smith Publication.

———. 1950. *I, Robot*. New York: Gnome Press. [Doubleday & Company, New York, 1950].

———. 1982. *Foundation's Edge*. New York: Doubleday.

———. 1993. *Forward the Foundation*. New York: Doubleday.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine Experiment. *Nature* 563: 59–64. [Crossref]

BBC. 2014. El robot que escribe noticias debuta en Los Ángeles, *BBC Mundo*, 19 March 2014. https://www.bbc.com/mundo/noticias/2014/03/140318_curiosidades_robot_periodista_la_times_az.

Bietti, Eletta. 2020. From Ethics Washing to Ethics Bashing: a View on Tech Ethics from Within Moral Philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020, pp. 210–219. Available at SSRN https://ssrn.com/abstract=3513182.

Boden, Margaret A. 2017. *Inteligencia Artificial*. Madrid: Turner Noema.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bowles, Samuel. 2016. *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*. New Haven: Yale University Press.

Burridge, Nicky. 2017. Artificial Intelligence Gets a Seat in the Boardroom. *Nikkei Asian Review.* https://asia.nikkei.com/Business/Artificial-intelligence-gets-a-seat-in-the-boardroom.

Calvo, Patrici. 2018. Ética de las cosas (EoT). Hacia una digitización socialmente responsable y moralmente válida del ámbito universitario. In *Un diseño universitario para la responsabilidad social*, ed. Alicia Andrés and Rosana Sanahuja. Castellón: Universitat Jaume I.

Chen, Xue-Wen, and Xiaotong Lin. 2014. Big Data Deep Learning: Challenges and Perspectives. *IEEE Access* 2: 514–525. https://ieeexplore.ieee.org/abstract/document/6817512.
[Crossref]

Conde, Aintzane, Ander Arriandiaga, J.A. Sanchez, Eva Portillo, S. Plaza, and I. Cabanes. 2018. High-Accuracy Wire Electrical Discharge Machining Using Artificial Neural Networks and Optimization Techniques. *Robotics and Computer-Integrated Manufacturing* 49: 24–38.
[Crossref]

Conill, Jesús. 2006. *Ética HermenÉutica. Crítica desde la Facticidad*. Madrid: Tecnos.

———. 2012. Neuroeconomía y neuromarketing: ¿más allá de la racionalidad maximizadora? In *Guía Comares de neurofilosofía práctica*, ed. Adela Cortina, 39–64. Granada: Comares.

———. 2019. *Intimidad corporal y persona humana. De Nietzsche a Ortega y Zubiri*. Madrid: Tecnos.

Conill and García-Marzá. 2012. Neuroeconomics and Neuromarketing. In *La comunicación empresarial y la gestión de los intnagibles en España y Latinoamérica*, ed. Justo Villafañe, 143–163. Madrid: Pearson.

Cortina, Adela. 1993. *Ética aplicada y democracia radical*. Madrid: Tecnos.

———. 2010. *Cordial Justice*. Madrid: Trotta.

Cotino, Lorenzo. 2017. Big Data and Artificial Intelligence. An Approach to Their Legal Treatment Based on Fundamental Rights. *Dilemata. Journal of Applied Ethics* 24: 131–150.

Dastin, Jeffrey. 2018. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G?feedType=RSS&feedName=topNews&utm_source=twitter&utm_medium=Social.

Donati, Pierpaolo. 2019. *Sociología relacional de lo humano*. Barañain: EUNSA.

European Commission. 2012. *Responsible Research and Innovation: Europe's Ability to Respond to*

Societal Challenges. Brussels. https://op.europa.eu/en/publication-detail/-/publication/2be36f74-b490-409e-bb60-12fd438100fe.

European Parliament. 2017. *Civil Law Rules on Robotics*. Brussels. https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html.

———. 2020. *Regulation of the European Parliament and of Council on European Data Governance (Data Governance Act)*. Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0767&from=en.

———. 2021. *Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, EUR-Lex - 52021PC0206. Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:e0649735-a372-11eb-9585-01aa75ed71a1.

Expansión. 2019. Un humano vence en debate a robot de IBM... que le dio batalla. https://expansion.mx/tecnologia/2019/02/12/un-humano-vence-en-debate-a-robot-de-ibm-que-le-dio-batalla.

Fresneda, Carlos, and Carlos Alba. 2018. Margaret Boden: La Inteligencia Artificial, como el hacha, se puede usar para el bien o para el mal. *El Mundo*. http://lab.elmundo.es/inteligencia-artificial/margaret-boden.html.

Gracia, Diego. 2001. La deliberación moral: el método de la ética clínica. *Medicina clínica (Barc)* 117: 18–23.

García-Marzá, Domingo. 2005. ¿Sentimientos virtuosos? El papel de los sentimientos en la vida moral. *Diálogo Filosófico* 62: 241–256.

———. 2016. Neuroética aplicada: las consecuencias prácticas del neuropositivismo. *Pensamiento* 72 (273): 881–900.

González-Esteban, E., and P. Calvo. 2022. Ethically Governing Artificial Intelligence in the Field of Scientific Research and Innovation. *Heliyon* 8 (2): e08946.
[Crossref]

Higham, Catherine F., and Desmond J. Higham. 2018. *Deep Learning: An Introduction for Applied Mathematicians*. https://arxiv.org/pdf/1801.05894.pdf.

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission. Available at https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

Hirasawa, Toshiaki, Kazuharu Aoyama, Tetsuya Tanimoto, Soichiro Ishihara, Satoki Shichijo, Tsuyoshi Ozawa, Tatsuya Ohnishi, Mitsuhiro Fujishiro, Keigo Matsuo, Junko Fujisaki, and Tomohiro Tada. 2018. Application of Artificial Intelligence Using a Convolutional Neural Network for Detecting Gastric Cancer in Endoscopic Images. *Gastric Cancer* 21 (4): 653–660. https://doi.org/10.1007/s10120-018-0793-2.
[Crossref]

Iba, Hitoshi. 2017. *Evolutionary Approach to Machine Learning and Deep Neural Networks*: *Neuro-Evolution and Gene Regulatory Networks*.

Ito, Joi. 2017. Designing our Complex Future with Machines. *Resisting Reduction: A Manifesto*. https://joi.

ito.com/weblog/2017/12/20/resisting-reduction.html.

———. 2019. Forget About Artificial Intelligence, Extended Intelligence is the Future. *Wired*. https://www.wired.co.uk/article/artificial-intelligenceextended-intelligence.

Jentzsch, Sophie, Patrick Schramowski, Constantin Rothkopf and Kristian Kersting. 2019. Semantics Derived Automatically from Language Corpora Contain Human-Like Moral Choices. In *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES*). Retrieved from https://ml-research.github.io/papers/jentzsch2019aies_moralChoiceMachine.pdf.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1: 389–399. https://doi.org/10.1038/s42256-019-0088-2. [Crossref]

Kearns, M., and A. Roth. 2020. *The Ethical Algorithm. The Science of Socially Aware Algorithm Desing*. Oxford: Oxford University Press.

L'Oreal. 2018. Artificial Intelligence to Recruit the Best Candidates for L'Oreal Spain. *L'Oreal*. 6 June 2018. http://www.loreal.es/periodistas/notas-de-prensa/2018/jul/inteligencia-artificial-para-reclutar-a-los-mejores-candidatos-de-loreal-espana.

Malle, Bertram F. 2016. Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots. *Ethics and Information Technology* 18: 243–256. [Crossref]

Marblestone, Adam H., Greg Wayne, and Konrad P. Kording. 2016. Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience* 10 (94): 1–41.

Marr, Bernard. 2016. *Big Data. La utilización del Big Data, el análisis y los parámetros Smart para tomar mejores decisiones y aumentar el rendimiento*. Teell.

Matsumoto, Tetsuzo. 2018. *The Day AI Becomes God. The Singularity will Save Humanity*. Cambridge (NZ): Media Tectonics.

Misuraca, Gianluca, and Colin van Noordt. 2020. *Science for Policy Report: AI Watch Artificial Intelligence in Public Services Overview of the Use and Impact of AI in Public Services in the EU*. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120399/jrc120399_misuraca-ai-watch_public-services_30062020_def.pdf.

Morozov, Evgeny. 2013. The Real Privacy Problem. *MIT Technology Review*. Consulted in https://www.technologyreview.com/s/520426/the-real-privacy-problem/.

Nguyen, Anh M. 2017. *AI Neuroscience: Visualizing and Understanding Deep Neural Networks*. Wyoming: University of Wyoming.

O'Neil, Cathy. 2016. *Weapons of Math Destruction How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publisher.

Pardo, Pablo. 2014, 9 de junio. Un algoritmo, sentado en el consejo de un fondo chino. *El Mundo*. https://www.elmundo.es/economia/2014/06/09/5394e0d722601df76f8b458f.html.

Rubio, Isabel. 2018. Amazon prescinde de una inteligencia artificial de reclutamiento por discriminar a las

mujeres. *El País*. 12 de octubre de 2018. https://elpais.com/tecnologia/2018/10/11/actualidad/1539278884_487716.html.

Sagristà, Anna. 2016. Algorithms: La bola de cristal en gestión de personal. *Talentier.* 16 December 2018. https://blog.talentier.com/algoritmos-en-gestion-de-personal.

SAM. 2018. *MAR Web.* 20 June 2018. http://www.politiciansam.nz.

Searle, John. 1984. *Mind, Brain and Science*. Cambridge: Harvard University Press.

———. 1988. *Mind, Brain, and Behaviour*. New York: Routledge.

Smith, John Maynard. 1982. *Evolution and the Theory of Game*. Cambridge: Cambridge University Press. [Crossref]

Stats NZ. 2018. Algorithm Assessment Report. New Zealand Government. New Zealand. https://data.govt.nz/use-data/analyse-data/government-algorithm-transparency.

Vidal, Macarena. 2018. China estrena presentadores artificiales de televisión. *El País.* 9 de noviembre de 2018. https://elpais.com/tecnologia/2018/11/09/actualidad/1541765605_369415.html.

Walker, Joseph. 2012. A la hora de contratar, las empresas sustituyen intuición por algoritmos. *The Wall Street Journal.* 21 September 2012. https://www.wsj.com/articles/SB10000872396390444032404578008900640503088.

Webb, Sarah. 2018. Deep Learning for Biology. *Nature* 554 (7693): 555–557. [Crossref]

Xerox. 2018. *Solución de automatización de flujos de trabajo de Xerox para el reclutamiento y la selección.* https://www.xerox.com/es-ni/servicios/automatizacion-de-procesos/reclutamiento-y-seleccion-de-personal.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_7

7. The Metaverse: Building a Digital Hyper-economy

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

> *So Hiro's not actually here at all. He's in a computer-generated universe that his computer is drawing onto his goggles and pumping into his earphones. Into the lingo, this imaginary place is kwon as the metaverse. Hiro spends a lot of time in the metaverse. It beasts the shit out of the U-Stor-It. (…) Like any place in Reality, the [metaverse] is subject to development. Developers can build their own small streets feeding off of the main one. They can build buildings, parks, signs, as well as things that do not exist in Reality, such as vast hovering overhead light shows, special neighborhoods where the rules of three-dimensional spacetime are ignored, and free combat zones where people can go to hunt and each other. The only difference is that since the [metaverse] does not really exist -it's just a computer graphics protocol written down of piece of paper somewhere- none of these things is being physically built. They are, rather, pieces of software, made available to the public over the worldwide fiber-optics network.*
>
> *Neal Stephenson, Snow Crash (1992, 24–25).*

Abstract
The aim of this chapter is to develop a deeper understanding of key features of the economy's digital transformation and likewise consider their impact. Today, emerging and intrinsically linked phenomena such as cryptocurrencies, non-fungible tokens (NFTs) and the metaverse are significantly altering the foundations of the economic system, the role of economic agents, the horizon for social action and modes of living in society more broadly. They are also redefining all other spheres of activity—such as politics, health, communication

and education—on the basis of economic values such as efficiency, efficacy or scarcity and, above all, techno-scientific values such as digital, virtual, neutral, objective, immediate, hyperconnected, hyper-datafied, hyper-automated, simultaneous and artificial.

Keywords Cryptoeconomics – NFT revolution – Metanomic ethic – Tokenization – Digital hypereconomics – Social pathologies

7.1 From the Digital Economy to the Artificial Hyper-economy

The advent of this new hyperdigitized world—characterized by a high degree of dependence on ICT (Floridi 2012), the conversion of social reality into computable data and metadata (Mayer-Schöenberger and Cukier 2013), the hybridization of the human and non-human in social relations (Donati 2019, 2023), the reduction of the human being to a mere terminal for the continuous flow of online data and metadata (Zuboff 2019) and, more recently, the pre-eminence of the virtual over the real (Calvo 2023)—have generated increasing complexity, and thereby imposed greater vulnerability on human beings (Conill 2019; García-Marzá 2020; Calvo 2020; Domingo-Moratalla 2013).

A key focus of this new context is the digital transformation of the economy. This process is generating an economic model founded on the design, application and use of smart digital technology across all the facets, areas and processes involved in its development. The outcome is commonly known as the *digital economy*, although this term can give rise to confusion due to its anachronism and its manifest incapacity to embrace the full range of idiosyncrasies encountered in the reality emerging from the diverse forms of designing, applying and using the technology derived from today's *datafication*. While the internet remains a fundamental mechanism for the twenty-first-century economy, its impact on this and other sectors cannot be explained merely in terms of the exchange of goods and services through computing devices.

The term *digital economy* was coined in 1995 by Don Tapscott in *The Digital Economy*: *Promise and Peril in the Age of Networked Intelligence* in order to define the emerging economy of production and consumption prompted by the emergence of the internet (Tapscott 1995: 5–6). However, today, while the internet remains a core element of today's economy, it is no longer possible, as it was in the past, to reduce the emerging economic reality of digital transformation to mere *digital connectivity*. For the economy, the emergence of

fields of applied knowledge and technological development, such as Big Data and Artificial intelligence, has allowed the *hyperconnectivity* of relevance (linking people, animals, things and processes into a hyperaesthetic network woven by the continuous mass flow of their data and behavioural metadata), the datafication of preference (converting economic agents' behaviour into objective, measurable, comparable and auditable data, and the algorithmization of decision-making) and the reduction of economic planning to artificially intelligent mathematical models.

The impact of all this is being felt in the productive, communicative, administrative and decision-making spheres and processes throughout the economic realm, and they strike a harsh blow at the foundations and fundamental tenets of the traditional economic model. An insightful concept for addressing these developments is sociologist Shoshana Zuboff's idea of *surveillance capitalism* as set out in her book *The Age of Surveillance Capitalism*: *The Fight for a Human Future at the New Frontier of Power* (2019). She discusses how it paves the way for the process and consolidation of an economic model characterized by the commodification of private and intimate data and metadata, the reduction of citizens to mere terminals for the mass flow of data and metadata, and their redefinition as a product of economic agents. It is, therefore, clearly difficult if not impossible) to reduce the responsibility and characteristics of the current economic moment to a process of *digital connectivity*, and the terms the *digital hyper-economy* or, simply, the *hyper-economy,* articulate the reality of the twenty-first century.

Hypereconomics can be understood as a term that refers to a new economic reality whose main characteristic is that it is hypertextual, represented by a "structured set of texts, graphics, and so on joined together by links and connections" (RAE 2018). The technological reduction of the human being to *homo data* opens up the possibility of them creating an avatar or virtual twin in cyberspace, so that economic agents can also become part of the set of *linkable things* which until recently was only possible for text and graphics.

However, the hyper-economy can also be understood as a new economic context whose main feature is dependence on ICTs. Seemingly this is currently the dominant trend of the twenty-first century economy, as without ICTs the economy could neither function nor would it be possible to incorporate, maintain and develop the agents it depends on. In other words, the resources or structures underpinning the current economic system are increasingly virtual and ICT-dependent, which makes it more and more difficult to access, remain a part of and empower oneself within the economy without using ICTs. This idea may be intrinsically linked to the principal context of the term *hyperhistory* coined by

Luciano Floridi in “Hyperhistory and the Philosophy of Information Policies” (2012).

However, in 2021, the digital transformation of the economic sphere took a surprising and unexpected turn with the emergence of crypto-art. Crypto-art is a subfield of cryptoeconomics, which is in turn a subfield of hypereconomics. It is based on blockchain technology and NFTs, and is attracting the attention of artists, collectors and enthusiasts for its ability to tokenize anything that can be sold as art in the digital market, in other words, for its capacity to convert any piece of digital art, with or without a reference in reality, into a scarce resource and an economic asset by encapsulating its market value (Calvo 2023).

The huge success of NFTs has led to the virtual gaining increasing value over the real and, with it, the possibility of recreating virtual worlds in which to host alternative lifestyles and establish an alternative economy. However, this growing trend (Fernández 2022) not only affects the art world. The supremacy of the virtual over the real is having a major impact on human relations and individual and communal forms of social life. It is affecting every sphere of activity. It is a virtual, anomic, progressive and highly addictive world where a person can be whoever they want and do whatever they want, whenever they want and wherever they want in relation to a *virtualized other*, be it a person or an algorithm. Hence, *hypereconomics* can also be understood in terms of the mathematical concept of hyperspace, as a hypothetical representation of a place beyond three-dimensional reality.

7.2 Cryptoeconomics: Ethical Aspects of the NFT Revolution

The irruption of *blockchain* technology has led to a revolution for the art world, the consequences of its unprecedented disruptive process are being strongly felt across all its areas and dimensions: economic, artistic, theoretical, aesthetic, reflective, sustainable, moral, etc. The emergence of so-called crypto-art or cryptographic digital art is of particular relevance. It is based on NFTs, smart contracts and cryptocurrencies which, according to their advocates, guarantees the ownership, authenticity, scarcity, exclusivity, immutability, verifiability and traceability of digital works.

Although there are precedents for this development, crypto-art was consolidated in 2017 through the trade in digital artworks. Until then, the problem with digital artworks was their high and rapid rate of replication across the network, which hindered the possibility of giving them a market value.

Scarcity as a dominant value of economics was extensively discussed by Carl Menger—founder of the Austrian School and one of the fathers of the marginalist revolution along with William Stanley Jevons and Leon Walras—in his seminal book *Principles of Economics* (1871). For Menger, economic goods are the source of individuals' personal progress in a society, as their consumption allows for optimal states of individual well-being. However, unlike non-economic goods, they are scarce and their appropriation is limited. Therefore, as they can never meet the existing demand—if they did, they would not be economic—agents must compete with each other to appropriate these goods and thereby satisfy their needs and desires, which is what underpins the development of these commodities. By focusing his study on the analysis of the scarcity of consumer goods, the appropriation of the best mechanisms to satisfy them, and the underlying competitiveness, Menger laid the foundations of the modern concept of economics, as "(…) the science which studies human behaviour as a relationship between given ends and scarce means which have alternative uses" (Robbins 1932, 15).

Using the blockchain technology that supports the *Ethereum* cryptocurrency —fungible, exchangeable *tokens*—digital artists were able to endow their works with an aura of scarcity, authenticity and proprietary control through *tokenization*; that is, by converting them into NFTs, which are non-exchangeable resources recognized, encrypted and protected by this cryptographic platform.

Tokenization is the process of turning anything into a scarce digital asset through the "(…) encapsulation of value in tradeable units of account, called tokens" (Freni et al. 2020). Thus, through the "(…) creation of a self-governed (tok)economic system, whose rules are programmed by the token designer" (Freni et al. 2020), tokenization offers the possibility of creating digital scarcity and, at the same time, eliminating the commercial intermediation that makes works of art more expensive, saving time and money in transactions and, therefore, lowering costs and improving profits.

From then on, the market value of digital works began to be established, and rose exponentially; on 11 March 2021 Mike Winkelmann, aka Beeple, sold his digital work *Everydays*: *The First 5000 Days* for 57 million euros (Fairfield and Trautman 2021; Joselit 2021; Dean 2021).

With the sale of Beeple's work, the trade in crypto-art or cryptographic digital art began to be seen as an investment-friendly field and its products considered economic assets, paving the way for the consolidation of the crypto-economy as a subset of the *hyper-economy*. In fact, the growing trade in crypto-art or cryptographic digital art has increasingly become an important subset of the so-called crypto-economy, which is itself a subset of the so-called *hyper-*

economy. In less than 4 years, between 23 June 2017 and 27 April 2021, crypto-art managed to generate 6.1 million trade transactions linked to the purchase and sale of 4.7 million NFTs, with a turnover of €787 million (Ante 2021; Nadini et al. 2021). Then, following the sale of Beeple's digital artwork sale, NFT-linked trading grew exponentially to a value at least $44.2 billion worth of cryptocurrency" (FY2021, 2) in cryptocurrencies linked to ERC-721 and ERC-1155 contracts, the two types of Ethereum smart contracts associated with NFT markets and collections (Chainalysis 2022).

Nevertheless, crypto-art has generated a range of opinions and mixed feelings among key figures in the art world: artists, academics, gallery owners, and so on. Three significant concerns stand out: it is a new financial bubble; it is a revolution; it is a failed idea with little scope for development.

Aside from the purely economic or artistic aspects of these three responses to *crypto-art* as a *crypto-economy*, a series of equally important ethical challenges demand consideration, such as the mistreatment or destruction of cultural heritage, the social consequences of speculation and the generation of economic bubbles, and finally the environmental unsustainability that, among other issues, is inherent to the world of NFTs at present.

On the one hand, the success of NFTs, based mainly on the repeated increase of their market value rather than on any intersubjective agreement about the creative, aesthetic or cultural value of the artwork in question, has given rise to trends and behavioural patterns of dubious legal and, above all, moral validity, for example, the mistreatment or destruction of artistic heritage. The purpose of such violence against cultural heritage is to transform the market value of a tangible artwork by converting it into an intangible NFT and then destroying the original. In this way, a sense of uniqueness, inviolability and virtual exclusivity is created, which increases the artwork's market value in each subsequent commercial transaction.

In this regard, a prescient example is the case of crypto-investors Injective Protocol, who converted the Banksy's painting *Morons* into an NFT and then set fire to it live, streaming the event to "(…) inspire tech enthusiasts and artists". Before being incinerated the painting was valued at 80,000 euros, while as an NFT it sold for 320,000 euros (García-Madrid 2021). In other words, what is now being sold is not any intrinsic quality of the work, but rather the certificate of authenticity for the digital copy.

As a result, this case has generated considerable debate about the legal and moral limits of NFTs. Among other reasons, this is because, as UNESCO stated in its 1972 Convention for the Protection of the World Cultural and Natural Heritage, it considers artistic heritage to be cultural property, and therefore

worthy of respect, care and protection by societies and their institutions and organizations (UNESCO 1972: Art. 1). In addition, the *UNESCO Culture for Development Indicators: Methodology Manual* (2014) stresses that heritage is very important for the culture and development of societies, "(…) as it constitutes the 'cultural capital' of contemporary societies. It contributes to the continual revalorization of cultures and identities and it is an important vehicle for the transmission of expertise, skills and knowledge between generations. It also provides inspiration for creativity and innovation, which result in contemporary and future cultural products". (UNESCO 2014: 135). Thus, as a universally recognized good, artworks, and cultural heritage more broadly, have an intrinsic, as well as extrinsic value that confers and demands respect, care and preservation. In other words, artworks have value regardless of what they can achieve—economic benefit, user experience, prestige, contemplative pleasure, and so on—so that their merely instrumental use erodes their sense of existence and promotes their decline and disappearance. Therefore, when an artwork is destroyed in order to increase its extrinsic value, namely its market price, the respect it deserves is undermined and the care and preservation it demands are violated.

On the other hand, the advocates of crypto-art promote NFTs as an unprecedented convention-breaking development that are generating something totally new, original and highly beneficial for all involved. For example, they offer a higher degree of democratization, control and profit within the art world. However, many of the criticisms of NFTs revolve around their reproduction, and even the exacerbation of long-standing problems that sustain the establish art market system while also hindering its development, such as speculation, misappropriation or the creation of financial bubbles.

Many experts, collectors and artists are concerned that the so-called NFT revolution is an economic bubble comparable to the *dotcom bubble* that occurred between 1995 and 2000, which led to a crisis across the sector, the collapse of businesses, redundancies and millions lost in the stock market value of the companies affected. As in the past, the incredible increase in the value of NFTs seems to be a response to major intervention in the art market by crypto-speculators, seeking to profit from the millions lost by naïve crypto-investors. There have recently been a number of symptomatic cases, such as the 2.5 million euros paid for the NFT of the screenshot of the first 24-character tweet sent by Jack Dorsey—CEO of Twitter—on the social network; the $560,000 paid for the NFT of an open access column written by Kevin Roose for *The New York Times*; the €612, 914 paid for the NFT of the *Nyan Cat* meme; the $1.5 million paid for *Quantum*, the first ever NFT; the €170,000 paid for the first text message of just

15 characters; the 57 million euros paid for the NFT of Beeple's work; the 450 million euros that the crypto-artist Ben Lewis hoped to earn by taking a JPG file of the painting *Salvador Mundi*—a work that some speculators insist on attributing without any scientific evidence to Leonardo Da Vinci—and modifying it minimally by placing a handful of dollars in the left hand, converting it into an NFT and renaming it *Salvador Metaverdi*; among many other cases (Frye 2021a).

The main concern with financial bubbles is that they recreate zero-sum games where those who promote them—a few—tend to keep the lion's share of the profits generated and leave the rest—especially the most vulnerable in the system—to foot the hyperbolic bill for the crisis generated. As Angus Deaton, winner of the 2015 Nobel Prize for Economics, declared, crises are designed to benefit the rich, because they make it easier for them to rewrite the rules. That is why education reform is not moving forward in the US, because the children of the rich go to private universities. The same goes for health. In crises, these things always get worse (Díaz 2012).

Finally, the NFT revolution conceals a deplorable characteristic overtly ignored by its most fervent supporters: the serious environmental footprint of its activity. Although significant progress has been made in recent years, these efforts are clearly insufficient. As David Malone and Karl J. O'Dwyer warned in 2014, when cryptocurrency systems were scarce and their activity certainly small compared to today, "(…) the power currently used for Bitcoin mining is comparable to Ireland's electricity consumption" (2014, 4). Today, crypto-art is an environmentally unsustainable system that contributes further uncertainty to the already worrying rate of climate change and its impact on the planet's natural and social ecosystem (de Vries 2018).

In this respect, the main problem with crypto-art is the high level of energy required for the *tokenization* of NFTs. In this regard, it is currently estimated that "(…) each single Ethereum transaction is estimated to cause 85.47 $kgCO_2$ resulting from the mining devices involved in verifying the transaction, and there were 942,812 NFT sales in the month preceding 10 October 2021" (Truby et al. 2022). In other words, a single NFT can generate many transactions—minting, bidding, ownership transfers, and so on—and the environmental impact of each transaction is equivalent to driving a petrol car for 1000 kilometres or flying a commercial airliner for 2 h. All this is due to the computational intensity of *Ethereum* 's *Proof of Work* (PoW) algorithm, which is used to ensure the system's inviolability and traceability (Akten 2020; Digiconomist 2022).

It should also be noted that NFT' energy use increases the inequality gap. It

has already been calculated that, in future, the environmental unsustainability of NFTs will lead to increased mortality among the most vulnerable groups in society.

> Furthermore, current emissions from mining devices supporting NFTs transactions are predicted to kill people at some time in the future. Bressler estimates that the average lifetime carbon emissions of 3.5 Americans (4434 metric tonnes or 4,434,000 kgCO2) will kill one person between 2020 and 2100 who would not otherwise have died. Death rates from blockchain transactions can then be estimated based on Bressler's calculations, by calculating the estimated emissions caused by a blockchain network and dividing the number of transactions to calculate an estimated emissions transaction cost (Truby et al. 2022).

There can be no doubt that the environmental impact of the use of different *blockchain* projects linked to NFTs is a major concern for society, as well as for the sector (Choobineh et al. 2022). It is seen as one of the major constraints to its success and wide-scale practical application.

> To mitigate the problems mentioned above, the community investigated alternative consensus mechanisms, based on more energy-efficient resources. One such consensus mechanism is Proof of Stake (PoS) that relies on the rationality of a stakeholder in the system to behave honestly due to the risk of devaluing the currency. In PoS, the consensus leader is chosen solely based on a function of their stake in the system (Deuber et al. 2020).

A clear signal of the sector's environmental concerns was *Ethereum*'s announcement to deactivate *Proof of Work* and adopt *Proof of Stake* in February 2022 (Mourya 2021; Fairley 2019). As stated on *Ethereum*'s corporate website, "As part of that roadmap, the existing proof-of-work chain (Eth1) would eventually be deprecated via the difficulty bomb. Users & applications would migrate to a new, proof-of-stake Ethereum chain, known as Eth2" (Ethereum.org 2022). Furthermore, other signs of change are the steady increase in proposals for new protocols that complement *Proof of Work*, such as *Proof of History*, or alternatives, such as *Proof of Hash Ownership*, *Two-Round PoW* and *Jakarta PoW* in order to reduce the environmental impact of the blockchain and enable its maximum deployment and potential in practice (Yakovenco 2019; Choobineh et al. 2022; Deuber et al. 2020; Chatbi 2021a, b, c).

Given the existing alternatives and the statement that it is feasible to design and implement increasingly robust and environmentally friendly protocols (David et al. 2018; Kiayias et al. 2017; Li et al. 2020; Tas et al. 2022; Yakovenco 2019), designers, implementers and the users of blockchain systems linked to NFTs can be expected to guide their behaviour and decision-making rationally by taking into account values such as sustainability, responsibility, fairness and respect for the environment. This is one of the key issues to building trust in the *blockchain*, and as a result, its wide-scale practical application.

Nevertheless, these actions are insufficient. The strong, rapid and ongoing viral spread of NFTs from the second half of 2021 onwards has encouraged investors and large technology corporations to explore how NFTs can be developed, applied and used in alternative contexts, which has led to an exponential increase in transactions linked to NFTs. Although individual transactions are more sustainable, the increasing number results in greater energy use, exacerbating the unsustainability of the blockchain protocols used for NFTs.

A clear example, as will be seen below, is the metaverse revolution. Only a few months after the aforementioned sale of *Everydays*: *The First 5000 Days*, Facebook founder and CEO Mark Zuckerberg presented a new business strategy based on the creation of interactive cryptovirtual worlds underpinned by blockchain technology, NFTs, cryptocurrencies such as Libra, and virtual reality (VR) tools such as the Cambria VR headset and Nazare VR goggles. The result of this has been the emergence of anomic, progredient and highly addictive environments capable of extracting the maximum development and potential from surveillance capitalism. It is a world where scarcity and market value will be controlled by artificially intelligent mathematical models and where, paradoxically, people will not only offer large companies privileged and unlimited access to their inner worlds but will also bear the costs of doing so.

7.3 The Metaverse: The Gateway to the Inner World

The metaverse, a concept coined by Neal Stephenson in his novel *Snow Crash* (1992) to refer to the world resulting from the future evolution of the internet, has erupted quite violently into the social and economic sphere following Mark Zuckerberg's announcement at the end of October 2021 that it was his business corporation's new strategic horizon. As a result, Facebook was renamed Meta to thereby reflect "(…) everything we do as, well as the future we want to help build" (Zuckerberg 2021: 1). While the corporation has continued to maintain its various applications to connect people through digital environments, such as Facebook and WhatsApp, according to Zuckerberg these are no longer the goal

of the company, which seeks to become the means to *bring the metaverse to life* (Zuckerberg 2021).

Based on virtual reality devices, such as computers, consoles, virtual reality glasses and headsets among other paraphernalia, blockchain technology, cryptocurrencies and NFTs, the metaverse aims to create a cryptovirtual, anomic, immersive, progredient and highly addictive world where people can engage in social and economic interaction with virtualized others—be they human or algorithmic (social and economic bots) avatars—while pretending to be whoever they want and aspiring to do whatever they want, whenever and wherever they want. The company claims that in the metaverse participants can go to the cinema to see the latest films; consult reproductions of the permanent and temporary collections of museums, as well as the digital collections of major libraries; play interactive games with thousands of virtual avatars of people or *bots* in the manner of the famous and lucrative Minecraft and Fortnite; attend fashion events and music concerts by major singers and bands; go to the doctor's surgery; graduate from a prestigious university (meta-university); find a romantic partner and go on a date; hold business meetings; buy, decorate and fill virtual homes with NFTs of digital art, designer furniture, designer clothes and all kinds of exclusive utensils; climb the 14 eight-thousander mountains; attend a political rally; or even help build the metaverse, and obtain benefits for doing so, among many other things. In short, the aim is to recreate a parallel virtual world and life where people can give free rein to their imagination.

Since the announcement made by the CEO of Facebook—now Meta—the creation of metaverses has continued to grow in number, content and business volume, giving rise to what has become known as metanomics or the metaverse economy. The most important platforms are Decentraland, The Sandbox, Axie Infinity, Somnium Space, Bloktopia, Start Atlas, Cryptovoxels and Uttopion. Many of these have existed for some years, but Meta's public presentation encouraged financial institutions and companies of all kinds to invest in this emerging sphere of business. Interest is such that during 2021, 268,645 plots of virtual land—digital assets represented by an NFT—were sold to companies and individuals with an average value of more than 11,500 euros. Companies such as Microsoft, Atari, Kiguel, Metaverse Group, Republic Realm have bought meta-towns in the Descentraland and Sandbox metaverses for a market value of up to €3.87 million, while owners of large companies such as Prestige Realty Group have bought up some of the nearly 9000 luxury meta-mansions offered for sale by the Keys metaverse at around €2200 euros each, as well as some of the 100 luxury meta-islands offered for sale by Sandbox, for which they have paid as much as $300,000. The motive behind these purchases is to use these virtual

spaces for business and work meetings among other purposes (Schenker 2022; Venkataramakrishnan and Steer 2022; the Financial Times 2022; van Rijmenam 2022). There is such a high degree of interest that the four main metaverses—Sandbox, Decentraland, Cryptovoxels and Somnium—achieved a total turnover of €451 million from this type of real estate trading during 2021, and, according to the *Metaverse Market Size, Share, Companies & Trends Analysis Report* prepared by Brand Essence Market Research (2022), "The Metaverse Market Size was valued at USD 209.77 Billion in 2021 and is expected to reach USD 716.5 Billion by 2027 with a CAGR of 22.7% over the forecast period" (Brand Essence Market Research (2022).

Companies designing and selling digital product have also engaged with this emerging market. The major clothing and fashion brands, such as Zara, Adidas, H&M, Gucci, Louis Vuitton, J.W. Anderson, Valentino or Marc Jacobs, for example, are already positioning themselves in the metaverse to launch their new clothing collections through meta-shops, and, above all, to launch lines of luxury NFT meta-products to personalize or customize *metaversian avatars*. According to data from Bain & Co., €50,000 million worth of customization products for video game avatars was sold in 2020. Likewise, Nike sold €3.5 million worth of NFT avatar shoes in just 5 minutes and The Dematerialized and Rebeca Minkoff sold a limited-edition collection of NFT handbags for 100 euros each in a matter of minutes. These are now sold on NFT sales websites for over 1000 euros. In short, as the report *Opportunities in the metaverse* drawn up by JPMorgan Chase Bank states:

> When you think about the economics of the metaverse—or metanomics—there are opportunities in almost every market area. Imagine you have an online avatar and you want to change what it/you are wearing, you will be able to buy limited-edition, digitally branded clothing that you pick after browsing a virtual showroom. Or you may start your own small business, such as an art gallery where you display your latest and greatest collections, or a virtual private club (JP Morgan Chase Bank 2022: 6).

Finally, the potential of the metaverse has also prompted interest from the sex industry. Firstly, with regard to the market for NFTs. Diversual, a sex toy design company, has produced the first collection of NFT vibrators for the metaverse. On the other hand, companies are seeking to explore the potential of haptic technology, which develops devices and interfaces that enable human-to-human interaction through the sense of touch. Owo and bHaptics, for example,

are two of the many companies currently developing and marketing haptic suits, jackets and gloves, an immersive technology that generates a second skin on the wearer through which they can feel different types of sensations, in different parts of the body and at different levels of intensity. These and other products are opening up new possibilities for cybersex and exponentially raising expectations about its commercial potential (Coloe 2021; Shanker and Zytko 2022; Zhao et al. 2022; Charamba 2022; Mystakidis 2022).

In short, the business possibilities the metaverse seemingly offers are enormous and extend to any economic sector, from finance to the products and services industries (digital assets). However, the development of the metaversian *(metanomic*) economic system is by no means wholly positive. It is beset by the same moral problems as underly the NFT revolution, such as speculation, environmental unsustainability and the destruction and mistreatment of cultural heritage, and this is exponentially increased by the *immersive* and *extensive* possibilities it offers. Furthermore, other concerns demand in-depth analysis. For example, the supremacy of the virtual over the real, the increase in the type and level of intensity of social and personal psychopathologies (alienation, personality alteration, addiction, promotion of immoral behaviour, etc.) and, above all, the increasing vulnerability caused by the impact derived from the metaverse's algorithms gaining privileged and unlimited access to individuals' private and intimate world of individuals: control of free will, heteronomy, dysfunction and/or inhibition of critical capacity, and so on.

7.4 Metanomic Ethics: When the Real Matters

The capacity of the emerging metaversian economic system—metanomics (JPMorgan 2022)—to exploit multiple ideas and technology in a virtual space and, above all, to generate interest among the financial markets and its users prompts misgivings among the public. Concerns are raised over its intended and unintended consequences that negatively affect markets and societies, especially among the most vulnerable groups.

With regard to the ethical concerns underpinning the design, application and use of metaversian frameworks, from an ethical-discursive perspective (this takes as its moral perspective those affected by a precept, action and/or decision, and it means that any rational justification is framed by the intersubjective acceptance of the consequences deriving from its present or future application) it is possible to discern several important issues. In addition to concerns about the creation of a new economic bubble that will eventually burst, with detrimental effects on society and the mistreatment of cultural heritage by those who seek

economic profit at any price, as well as the energy use involved in maintaining and developing the metaverse, people are also worried about what impact an increasing disinterest in the real world will have; the psychosocial pathologies the metaverse's distortion of reality will give rise to, the consequences of the privileged access to the private and intimate world that *meta-users* uncritically grant to the algorithms of metaversian systems; and, finally, how the creation of alternative realities will make injustice invisible. In other words:

– *The supremacy of the virtual over the real:*

The broader value of a *meta-user's* experience, whether selling or buying a virtual asset, enjoying an evening out with their partner or friend, attending a concert, visiting a museum exhibition, conducting a business deal or offering a thought-through opinion on a topical issue, is beginning to be perceived as far superior to experiences based on direct, sentient, relational encounters with people. The aforementioned examples of the destruction and mistreatment of cultural heritage provide a clear example of this, but they are not the only ones. People's interest and attachment to immersive virtuality are increasing: visits to virtual museums, online dating for partners and sexual relations (Shanker and Zytko 2022), the purchase of all kinds of NFT digital assets—meta-islands, meta-yachts and meta-mansions for business meetings or leisure—as well as and the purchase and *customization* of avatars, virtual tourism, participation in virtual opinion spaces, and many other forms of engagement.

– *The increasing number and intensity of psychosocial pathologies*

This supremacy of the real over the virtual produces an increase in the type and level of intensity of social and personal psychopathologies (Siurana 2021, 2023; Domingo Moratalla 2013), such as alienation, personality disorders, addiction, social dysfunction, depression and anxiety, aggression and lack of self-control, lower interpersonal skills, social anxiety, and increased feelings of loneliness. As Han, Bergs & Moorhouse state (2022: 13), "(…) the limitations of the proposed future research streams should be acknowledged, we believe that negative psychological consequences of VR consumer experience escapes in the metaverse will affect social interactions as well as consumers' physical and psychological well-being. Inevitably, it will affect peoples' ability to cope and function in life".

Obsessive shopping is an important addiction, especially the growing schizophrenic phenomenon of buying and customizing avatars with all kinds of NFT products, such as branded clothing and luxury goods. As reported by Reuters (Howcroft 2022), one famous model –Imani McEwan– spent more than

15,000 euros in 2 months on NFT clothing and products for her metaversian avatar (berets, jumpers, T-shirts, and so on), because, she believes, “Your avatar represents you. Basically, what you wear is what makes you who you are”. The addictive power of the metaverse is so great that major fashion houses such as Louis Vuitton, Burberry and H&M have created lines of luxury NFT clothing and accessories for metaversian avatars.

Alienation is defined by Hartmut Rosa as “(…) a specific form of relationship to the world in which the subject and the world confront each other with indifference or hostility” (Rosa 2016). Virtual and immersive environments such as the metaverse generate an alteration and/or loss of reason and meaning, while also leading to a deformation of the personality and identity of a person or group. The forces giving rise to this field of distortion of social reality and human behaviour are, firstly, the hybridization between the human and the (ro)bot, and between the real and the virtual in social relations. Secondly, there are the processes of artificial polarization, which, forged from different expressive uses of communication and “(…) motivated by the search for status or emotional release—mainly on the internet—form a false impression of political polarization or indignation about an issue” (Pérez-Zafrilla 2021: 3). Finally, there is the phenomenon of social acceleration; structured around three basic categories—technological acceleration, the acceleration of social change and the acceleration of the pace of life—giving rise to the desynchronization of the different functional spheres (López-González 2022). The metaverse promotes deep mental states in metanomic agents, which significantly alter an individual’s capacity to exercise free will—autonomy, critical judgement, moral sentiments and prosocial emotions. This leads to deficits of empathy, high levels of inaction in the face of injustice, emotional harm, moral disorientation, diminished cognitive performance, and disinterest in critiquing conventional social customs and their consequences, among many other issues (Bucci et al. 2019; Han et al. 2022; López-González 2022; Wibirama et al. 2020; Lavoie et al. 2020; Mittelstaedt et al. 2020; Nichols and Patel 2002; Sherman and Judkins 1992; Yee et al. 2009 Hillier and Sethi (2013).

– *Privileged access to people’s inner worlds*

For the algorithms that govern metaverse environments, these virtual spaces constitute a privileged and almost unlimited gateway to the private and intimate world of metanomic agents (Conill 2019; Domingo Moratalla 2022; Siurana 2023; Pérez-Zafrilla 2021). The immersive devices and technologies used, such as helmets, virtual reality glasses and haptic suits, allow the collection of sensory and even neural data from meta-users (Bermejo and Hui 2022; Hillmann 2021).

As a result *meta-users'* capacity to control their own free will is put at risk. Furthermore, the sensory data that they and other *meta-users* provide to *metaversian* algorithms and their collective processing can serve to detect patterns, anomalies, flaws and behavioural weaknesses in society as a whole, as well as in specific individuals. As a result, these individual and collective weaknesses can be used by algorithms to deform perceptions of reality in order to achieve a strategic purpose by promoting heteronomy and manipulating the individuals' free will and the common will of society through the use of echo chambers, bubble filters, artificial polarization, and so on.

– *The prevalence and promotion of immoral patterns and behaviours*

A lack of rules is one of the keys to the success of the metaverse. In these environments, *meta-users* can be whoever they want, and behave however they want, whenever they want and wherever they want. Greater enforcement of precepts, therefore, becomes less attractive to *meta-users*, who head for other spaces where they find forms of behavioural anomie more akin to their preferences. However, the anomie that governs metaverse environments encourages behaviour that is homophobic, xenophobic, misogynistic, aporophobic and so on, together with a reduction in self-control, while promoting socially and morally reprehensible and criminal patterns of conduct both inside and outside the metaverse, such as the occupation of spaces, the reproduction and sale of objects and brands, and even sexual harassment or gang rape using avatars (Bermejo and Hui 2022; Schwirn 2022).

The Meta Corporation's metaversal platform Horizon Worlds has already recorded two sexual attacks on avatars with feminized characteristics. One of them, the avatar of Nina Jane Patel, was harassed and raped by four avatars with masculine features just 60 seconds after logging on (Oppenheim 2022). As the victim herself explained in a press release, the sensation was traumatic, since: "Virtual reality has essentially been designed so the mind and body can't differentiate virtual/digital experiences from real. In some capacity, my physiological and psychological response was as though it happened in reality" (Patel 2021: 1). On the other hand, the virtual environment Next Earth, intended for digitizing cities and selling NFTs of their houses and plots, has received complaints from the owners of the real plots in these cities for having sold them to users without their permission. Although buying a digitized version of real house in the metaverse has no legal significance in terms of ownership, it may have collateral significance in terms of image, publicity or competition (Okonkwo 2021; Schwirn 2022).

– *The construction of an alternative reality:*

The metaverse can also be seen as the application of an invisibility cloak for clearly unjust situations. It permits the construction of alternative worlds where there are no limits to illegal and immoral behaviour; a parallel world where only those with sufficient money can enter, and an exclusionary world where justice and fairness are "invented" to satisfy our preferences, and where all those who do not have something to offer are barred from entry, and therefore from existence.

In short, whatever the current *tokenization of the economy* has become, it is already happening, and it is undoubtedly one of the great challenges for the economy and democracy in the twenty-first century. The emergence of the metaverse is having a disruptive impact on the proper development of functional spheres such as the economy and politics, as well as society and citizenship. Among other issues, this is due to it subtly deforming rationalization processes with regard to the meaning of a range of things in this world, as well as markedly altering people's personality and character. It is therefore necessary to identify a set of principles for an ethically designed metaverse, as well as to draw up, implement and apply ethical governance systems across the various metaverses to help mitigate or eradicate their negative consequences for the economy, the police, society and individuals. In other words, what is needed is metanomic ethics as applied ethics, capable of criticizing and orienting metaversian systems in a just and *pleasing* way.

By taking this ethical approach, an attempt can be made to unravel the *normative assumptions* underlying the metaverse—the frame of reference—based on which it is possible to rationally criticize, orient and motivate both the goals and aims pursued by those who create these spaces, and the behaviour of those who use them. It can also help design and implement methods for resolving moral problems and conflicts, as well as identifying and promoting the civic virtues that enable an exemplary character to be upheld in every sphere of metaversal activity and relationships. Finally, it can also help to identify the tools, mechanisms and guidelines that allow the practical application and fact-based re-creation of metaverses, such as codes, guidelines, ethical committees and audits, and explicability reports.

In this respect, and as will be seen over the course of the chapters that follow, an approach to ethics such as that proposed by Apel and Habermas, yet whose scope has been expanded through the introduction of a necessary axiological, sentient, experiential, virtuous and applied moment in order to foster its practical development (Cortina et al. 2003; Cortina 1986, 2007, 2010, 2017; Conill 2006,

2019; García-Marzá 1992, 2019; Siurana 2003), can prove fruitful when facing the ethical challenges of the metaverse.

This discursive ethical approach engages with intersubjectivity in order to provide a criterion of rationality from which it is possible to criticize the design, implementation and use of metaverse environments: the intersubjective consensus given by all those affected in the present and in the future, on the acceptance of the consequences of a precept, action or decision that is applied or enforceable.

However, by taking a broader and more mature approach in the form of the introduction of cordial, experiential, sentient, axiological and prudent resignation enables us to tackle its application across a range of functional spheres (Cortina 1986; Conill 2006; García-Marzá 1992), such as metanomics. In other words, as well as being concerned with recognizing the moral validity of universally binding precepts that govern the functional and relational spaces occupied by free and communicatively connected beings, this approach also has the potential to address concerns about the principles, values, virtues, ends, motives for action and effects that orient, cohere and motivate the behaviour of communicatively and affectively connected agents.

In short, the emergence of the metaverse has become one of the most disruptive elements affecting the way modern societies and their different spheres—in particular the economy—function and develop. There seems little doubt that for the economy, the creation of alternative virtual worlds underpinned by algorithms offers great possibilities for obtaining and maximizing profits in the short, medium and long term. Yet there can be no question that their uncritical and unethical design, application and use can be highly detrimental to the economy itself and, above all, to society, especially the most vulnerable groups, who are excluded, denying their very existence.

As this chapter has shown, the uncritical and unethical design, implementation and use of metaverses can have highly damaging impacts on the digitally hyperconnected society. In this respect, the environmental unsustainability of the metaverse system has been highlighted; the rise and increasing intensity of psychosocial pathologies among meta-users; greater vulnerability and alienation due to people’s increasing exposure to the intelligent algorithms that control the metaverse; and, the longevity and promotion of undesirable models and forms of behaviour for a morally mature society.

The consequences of this *hyperdigitized economy* being designed and promoted by a range of public and private organizations are revealed in the exponential increase in spaces in the lifeworld in which objective reality undergoes distortion in addition to an emergence of anomie, psychosocial

pathologies and a decline in meaningful activity in the space where people acquire the convictions and self-evidence necessary to understand aspects of the world (Habermas 1987). By contrast to the virtual, the value ascribed to the real is declining with astonishing speed, while economic agents are becoming engulfed in a virtual world of data and behavioural metadata leaving them unable to discern what role they should perform: whether they represent the aim to be pursued or are the product of it. In short, has this virtual world (the metaverse) and its instruments (cryptocurrencies, blockchain, NFTs and virtual reality technologies, among others) been created to increase their capacity to satisfy their interests and desires—even their deepest and most unspeakable urges? Or in order to gather data, including the most intimate information, about people's tendencies and behaviour in order to control their impulses? The hyperdigitized economy therefore seems to be launching the conquest of its final frontier: access to the economic agent's inner world.

To address these and other ethical challenges underlying the design, implementation and use of the metaverse, it has been proposed that there is a need to develop a *metanomic ethics* as an applied ethics which would enable a critique of metaversian environments and guide them in a direction that would be just and pleasing for all parties involved and affected. This would be, in short, discursive digital ethics, whose proposed clarification, foundation and application of moral knowledge, as will be seen in the chapters that make up the third and final section of this book, can establish a solid basis for framing or integrating algorithmic democracy into a deliberative democracy.

Bibliography

Akten, M. 2020. The Unreasonable Ecological Cost of Cryptoart (Part 1). *Medium*. Available at: https://memoakten.medium.com/the-unreasonable-ecological-cost-of-cryptoart-2221d3eb2053.

Ante, L. 2021. The NFT Market and Its Relationship with Bitcoin and Ethereum. *BRL Working Paper Series No. 20.* Available at SSRN: https://ssrn.com/abstract=3861106 or https://doi.org/10.2139/ssrn.3861106.

Atlanta. Retrieved from https://arxiv.org/abs/1706.07068v1.

Bermejo, C., and P. Hui. 2022. Life, the Metaverse and Everything: An Overview of Privacy, Ethics, and Governance in Metaverse.

Brand Essence Market Research. 2022. Metaverse Market Size, Share, Companies & Trends Analysis Report [BMRC 1740]. brandessenceresearch.com. https://brandessenceresearch.com/technology-and-media/metaverse-market-size.

Bucci, S., Matthias Schwannauer, and Natalie Berry. 2019. The Digital Revolution and Its Impact on Mental

Health Care. *Psychol Psychother* 92 (2): 277–297. https://doi.org/10.1111/papt.12222. Epub 2019 Mar 28. [Crossref]

Calvo, Patrici. 2020. The Ethics of Smart City (EoSC): Moral Implications of Hyperconnectivity, Algorithmization and the Datafication of Urban Digital Society. *Ethics and Information Technology* 22: 141–149. https://doi.org/10.1007/s10676-019-09523-0.
[Crossref]

———. 2023. Cryptoart: Ethical Challenges of the NFT revolution. *Ethics and Information Technology* (in press).

Case, Anne, and Angus Deaton. 2020. *Deaths of Despair and the Future of Capitalism*. Princeton: Princeton University Press.
[Crossref]

Chainalysis. 2022. *The Chainalysis 2021 NFT Market Report.* https://blog.chainalysis.com.

Charamba, K. 2022. *Beyond the Corporate Responsibility to Respect in the Dawn of a Metaverse*. University of Hong Kong Faculty of Law Research Paper No. 2022/14. Available at SSRN: https://ssrn.com/abstract=4043254 or https://doi.org/10.2139/ssrn.4043254.

Chatbi, Charaf E 2021a. *Proof of Hash Ownership*. https://www.researchgate.net/publication/356554200.

Chatbi, Charaf E. 2021b. *JPoW and TRPoW Variants of Proof of Work*. https://www.researchgate.net/publication/356414880.

———. (2021c). *Variants of Proof of Work.* https://www.researchgate.net/publication/356554200.

Choobineh, M., A. Arab, A. Khodaei, and A. Paaso. 2022. Energy Innovations Through Blockchain: Challenges, Opportunities, and the Road Ahead. *The Electricity Journal* 35 (1): 107059. https://doi.org/10.1016/j.tej.2021.107059.
[Crossref]

CoinMarketCap. 2022. *Today's Cryptocurrency Prices by Market Cap.* https://coinmarketcap.com.

Coloe, D.S. 2021. The Metaverse Is Coming. *Patheos*. https://www.patheos.com/blogs/newgenerationsexplorefaith/2021/12/the-metaverse-is-coming/.

Conill, Jesús. 2006. *Ética HermenÉutica. Crítica desde la Facticidad*. Madrid: Tecnos.

Conill, J. 2019. *Intimidad corporal y persona humana. De Nietzsche a Ortega y Zubiri*. Madrid: Tecnos.

Cortina, Adela. 1986. *Ética mínima*. Madrid: Tecnos.

———. 2007. *Ética de la razón cordial. Educar en la ciudadanía en el siglo XXI*. Oviedo: Nobel.

———. 2010. *Justicia Cordial*. Madrid: Trotta.

———. 2017. *Aporofobia, el rechazo al pobre*. Barcelona: Paidós.

———. 2019. Ética de la inteligencia artificial. *Anales de la Real Academia de Ciencias Morales y Políticas* 96: 379–394.

Cortina, Adela, Domingo García-Marzá, and Jesús Conill. 2003. *Public Reason and Applied Ethics. The Ways of Practical Reason in a Pluralist Society*. New York: Routledge.

David, B., P. Gaži, A. Kiayias, and A. Russell. 2018. Ouroboros Praos: An Adaptively-Secure, Semi-synchronous Proof-of-Stake Blockchain. In *Advances in Cryptology – EUROCRYPT 2018, Lecture Notes in Computer Science*, ed. J. Nielsen and V. Rijmen, vol. 10821, 66–98. Cham: Springer. https://doi.org/10.1007/978-3-319-78375-8_3.
[Crossref]

Dean, Sam. 2021. $69 Million for Digital Art? The NFT Craze, Explained. *Los Angeles Times.* https://www.latimes.com/business/technology/story/2021-03-11/nft-explainer-crypto-trading-collectible. Accessed 2021.

de Vries, Alex. 2018. Bitcoin's Growing Energy Problem. *Joule* 2 (5): 801–805.
[Crossref]

Deuber, D., N. Döttling, B. Magri, G. Malavolta, and S.A.K. Thyagaraja. 2020. Minting Mechanism for Proof of Stake Blockchains. In *Applied Cryptography and Network Security*, ed. Conti, Mauro, Zhou, Jianying, Casalicchio, Emiliano and Spognardi, Angelo, 18th international conference ACNS 2020, Rome, Italy, 19–22 October, 2020, proceedings, part I, pp. 315–334.

Díaz, I. 2012. Angus Deaton: Las Crisis están creadas para beneficiar a los ricos. *El País, XL Semanal*.

Digiconomist. 2022. *Bitcoin Energy Consumption Index.* https://digiconomist.net/bitcoin-energy-consumption/.

Domingo-Moratalla, Agustín. 2013. *Educación y redes sociales: la autoridad de educar en la era digital*. Madrid: Encuentro.

———. 2022. Hermenéuticas de la intimidad: horizontes para educar en una sociedad hiperconectada. *Estudios Filosóficos* LXXI: 95–109.

Donati, P. 2019. *Sociología relacional de lo humano*. Barañain: EUNSA.

Donati, Pierpaolo. 2023. The Prediction of Social Catastrophes: Between Necessity and Contingency. *Journal for the Theory of Social Behaviour* 1–21. https://onlinelibrary.wiley.com/doi/10.1111/jtsb.12399. https://doi.org/10.1111/jtsb.12399.

Ethereum.org (2022, 24 January). The Great Renaming: What Happened to Eth2?. *Ethereum.org*. https://blog.ethereum.org/2022/01/24/the-great-eth2-renaming/.

Fairfield, Joshua, and Lawrence J. Trautman. 2021. Virtual Art and Non-fungible Tokens. *SSRN*: 1–71.

Fairley, P. 2019. Ethereum Plans to Cut Its Absurd Energy Consumption by 99 Percent. *IEEE Spectrum* https://spectrum.ieee.org/ethereum-plans-to-cut-its-absurd-energy-consumption-by-99-percent.

Fernández, P. 2022. Rafa Nadal will launch his own NFT collection with autograph and the networks light up. *Cardaniers.* https://cardaniers.com/en/rafa-nadal-lanzara-su-propia-coleccion-nft-con-autograph-y-las-redes-se-encienden/?amp.

Floridi, L. 2012. Hyperhistory and the Philosophy of Information Policies. *Philosophy & Technology* 25: 129–131. https://doi.org/10.1007/s13347-012-0077-4.
[Crossref]

Freni, P., E. Ferro, and R. Moncada. 2020. Tokenization and Blockchain Tokens Classification: A Morphological Framework. *IEEE Symposium on Computers and Communications (ISCC)* 1–6. https://doi.org/10.1109/ISCC50000.2020.9219709.

Frye, B.L. 2021a. NFTs & the Death of Art. *SSRN*. Available at https://ssrn.com/abstract=3829399 or https://doi.org/10.2139/ssrn.3829399.

———. 2021b. Adopt a Design Patent: 1842–1925. *SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3934074.

García-Madrid, F. 2021, 20 March. Virtudes y peligros del criptoarte que enloquece a los mercados. *La Vanguardia*. Retrieved from https://www.lavanguardia.com/cultura/20210320/6500213/criptoarte-virtudes-peligros-mercados-banksy.html.

García-Marzá, Domingo. 1992. *Ética de la justicia: J. Habermas and discursive ethics*. Madrid: Tecnos.

———. 2016. Neuroética aplicada: las consecuencias prácticas del neuropositivismo. *Pensamiento. Revista de Investigación e Innovación Tecnológica* 72 (273): 881–900.

———. 2019. Ética y democracia. Notas para una renovada ética del discurso. In *Ética y democracia desde la razón cordial*, ed. Elsa González, Juan Carlos Siurana, José Luis López, and Marina García-Granero, 7–16. Granada: Comares.

———. 2020. Against the New Democratic Revisionism. *Ethics* 19 (3): 568–591. [Crossref]

Habermas, J. 1987. *Teoría de la acción comunicativa II: Crítica de la razón funcionalista*. Madrid: Taurus.

Han, D.I.D., Y. Bergs, and N. Moorhouse. 2022. Virtual Reality Consumer Experience Escapes: Preparing for the Metaverse. *Virtual Reality* 26: 1443–1458. https://doi.org/10.1007/s10055-022-00641-7. [Crossref]

Hillier, B., and F. Sethi. 2013. Two Case Reports: 'Delusions' Arising in Virtual Reality. *Journal of Psychiatric Intensive Care* 10 (1): 51–56. https://doi.org/10.1017/S1742646413000174. [Crossref]

Hillmann, Cornel. 2021. Practical Approaches: UX and XR in the Real World. In *UX for XR*. Berkeley: Design Thinking Apress. https://doi.org/10.1007/978-1-4842-7020-2_6. [Crossref]

Howcroft, Elizabeth. 2021, 12 August. Crypto Fashion: Why People Pay Real Money for Virtual Clothes. *Reuters.* https://www.reuters.com/business/finance/crypto-fashion-why-people-pay-real-money-virtual-clothes-2021-08-12/.

———. 2022. Crypto Fashion: Why People Pay Real Money for Virtual Clothes. *Reuters*. https://www.reuters.com/business/finance/crypto-fashion-why-people-pay-real-money-virtual-clothes-2021-08-12/.

Joselit, David. 2021 NFTs, or the Readymade Reversed, *October*, (175): 3–4 https://doi.org/10.1162/octo_a_00419.

JPMorgan. 2022. *Opportunities in the Metaverse*. JPMorgan Chase Bank. https://www.jpmorgan.com/

content/dam/jpm/treasury-services/documents/opportunities-in-the-metaverse.pdf.

Kiayias, A., A. Russell, B. David, and R. Oliynykov. 2017. Ouroboros: A Provably Secure Proof-of-Stake Blockchain Protocol. In *CRYPTO 2017. LNCS*, ed. J. Katz and H. Shacham, 357–388. Cham: Springer. https://doi.org/10.1007/978-3-319-63688-7_siu12.
[Crossref]

Lavoie, R., K. Main, C. King, and D. King. 2020. Virtual Experience, Real Consequences: The Potential Negative Emotional Consequences of Virtual Reality Gameplay. *Virtual Reality* 25: 69–81.
[Crossref]

Li, A., X. Wei, and Z. He. 2020. Robust Proof of Stake: A New Consensus Protocol for Sustainable Blockchain Systems. *Sustainability* 12 (7): 2824. https://doi.org/10.3390/su12072824.
[Crossref]

López-González, J.L. 2022. *La ética ante la cinética del turismo. Aportaciones desde la teoría crítica de la resonancia de Hartmut Rosa*. Castellón: Universitat Jaume I.

Malone, D., and K.J. O'Dwyer. 2014. Bitcoin Mining and Its Energy Footprint. In *Proceeding of the 25th IET Irish Signals System Conference (ISSC'14)*, pp. 280–285.

Mayer-Schöenberger, V., and K.N. Cukier. 2013. *Big Data. A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

Menger, C. 1871. *Grundsätze der Volkswirtschaftslehre*. Vienna: Wilhelm Braumüller.

Mittelstaedt, J.M., J. Wacker, and D. Stelling. 2020. VR Aftereffect and the Relation of Cybersickness and Cognitive Performance. *Virtual Reality* 23: 143–154.
[Crossref]

Mourya, Ekta. 2021, 18 October. Ethereum Developers Plan to Switch Off Proof-of-Work Forever by February 2022. *FXstreet* https://www.fxstreet.com/cryptocurrencies/news/ethereum-developers-plan-to-switch-off-proof-of-work-forever-by-february-2022-202110180748.

Mystakidis, Stylianos. 2022. Metaverse Mystakidis, Stylianos. 2022. "Metaverse". *Encyclopedia* 2 (1): 486–497. https://doi.org/10.3390/encyclopedia2010031.
[Crossref]

Nadini, Matthieu, Laura Alessandretti, Flavio Di Giacinto, Mauro Martino, Luca Maria Aiello, and Andrea Baronchelli. 2021. Mapping the NFT Revolution: Market Trends, Trade Networks and Visual Features, *arXiv*. Available at: arXiv:2106.00647v2 [q-fin.ST].

Nichols, S., and H. Patel. 2002. Health and Safety Implications of Virtual Reality: A Review of Empirical Evidence. *Applied Ergonomics* 33 (3): 251–271.
[Crossref]

Okonkwo, Ifeanyi E. 2021. NFT, Copyright; and Intellectual Property Commercialisation. Available at SSRN: https://ssrn.com/abstract=3856154 or https://doi.org/10.2139/ssrn.3856154.

Oppenheim, Maya. 2022. Woman Reveals 'Nightmare' of Being 'Gang Raped' in Virtual Reality". *The Independent.* https://www.independent.co.uk/news/uk/home-news/metaverse-gang-rape-virtual-world-b2005959.html.

Patel, Nina Jane. 2021. Reality or Fiction?. *Kabuni*. https://medium.com/kabuni/fiction-vs-non-fiction-98aa0098f3b0.

Pérez-Zafrilla, P.J. 2021. Artificial Polarisation: How Expressive Discourses Inflame the Perception of Political Polarisation on the Internet. RECERCA. Revista De Pensament I Anàlisi, 26(2). https://doi.org/10.6035/recerca.4661.

Plato. 1992. *Republic*. Madrid: Ed. Gredos.

———. 2000. *The Republic* [translated by tom Griffith, and edited by Giovanni R. F. Ferrari]. Cambridge: Cambridge University Press.

RAE. 2018. Hipertextualidad. *Diccionario de la Lengua Española*. Retrieved from https://dle.rae.es/hipertexto.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: MacMillan&co.

Rosa, Hartmut. 2016. *Alienation and Acceleration: Towards a Critical Theory of Temporality in Late Modernity*. Madrid: Katz.

Schenker, Jason. 2022. *Metaconomics: The Future of Business in the Metaverse Economy*. Prestige Professional Publishing.

Schwirn, Martin. 2022. A Legal Minefield Called the Metaverse. *Computer Weekly*. https://www.computerweekly.com/feature/A-legal-minefield-called-the-metaverse.

Shanker, Sarath S., and Douglas Zytko. 2022. The…Tinderverse?: Opportunities and Challenges for User Safety in Extended Reality (XR) Dating Apps. *ArXiv*. arXiv:2203.15120.

Sherman, B., and P. Judkins. 1992. *Glimpses of Heaven, Visions of Hell: Virtual Reality and Its Implications*. London: Hodder & Stoughton.

Siurana, Juan Carlos. 2003. *Una brújula para la vida moral: la idea de sujeto en la ética del discurso de Karl-Otto Apel*. Comares: Granada.

———. 2021. *Ética Para influencers* [ethics for influencers]. Madrid: Plaza y ValdÉs.

———. 2023. Biohermenéutica crítica para evaluar el impacto de los videojuegos en la naturaleza humana. In *Hermenéutica crítica y razón práctica: Homenaje a Jesús Conill*, ed. Domingo García Marzá, Juan-Antonio Nicolás Marín, and Agustín Domingo Moratalla, 473–482.

Stephenson, Neal. 1992. *Snow Crash*. New York: Bantam.

Tapscott, Don. 1995. *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*. McGraw-Hill Education.

Tas, Ertem Nusret, Fisher Yu, David Tse, and Sreeram Kannan. 2022. Reusing Bitcoin Mining to Enhance Proof-of-Stake Security. arXiv:2201.07946v1 [cs.CR] 20 Jan 2022. https://arxiv.org/pdf/2201.07946.pdf.

The Financial Times. 2022. NFTs: The Metaverse Economy. *Crypto.com – The Financial Times.* https://www.ft.com/partnercontent/crypto-com/nfts-the-metaverse-economy.html.

Truby, Jon, Rafael Dean Brown, Andrew Dahdal, and Imad Ibrahim. 2022. Blockchain, Climate Damage, and Death: Policy Interventions to Reduce the Carbon Emissions, Mortality, and Net-Zero Implications of Non-fungible Tokens and Bitcoin. *Energy Research & Social Science* 88: 1–14.
[Crossref]

UNESCO. 1972. *Convención sobre la protección del patrimonio mundial, cultural y natural*. Paris, UN. https://whc.unesco.org/archive/convention-es.pdf.

———. 2014. Heritage. In *UNESCO Culture for Development Indicators*. Paris, UN. https://en.unesco.org/creativity/sites/creativity/files/cdis_methodology_manual_0_0.pdf.

van Rijmenam, Mark. 2022. *Step into the Metaverse: How the Immersive Internet Will Unlock a Trillion-Dollar Social Economy*. Willey.

Venkataramakrishnan, Siddharth, and George Steer. 2022. Would You Buy a Home in the Metaverse? *The Financial Times*. https://www.ft.com/content/f5d5bed3-8cf2-40c6-8f85-6b824a223ffe.

Wibirama, S., P.I. Santosa, P. Widyarani, N. Brillanto, and W. Hafidh. 2020. Physical Discomfort and Eye Movements During Arbitrary and Optical Flow-Like Motions in Stereo 3D Contents. *Virtual Reality* 24: 39–35.
[Crossref]

Yakovenco, Anatoly. 2019. How Solana's Proof of History Is a Huge Advancement for Block Time. *Medium*, https://medium.com/solana-labs/how-solanas-proof-of-history-is-a-huge-advancement-for-block-time-178899c89723.

Yee, N., J.N. Bailenson, and N. Ducheneaut. 2009. The Proteus Effect: Implications of Transformed Digital Self-Representation on Online and Offline Behavior. *Communication Research* 36 (2): 285–312. https://doi.org/10.1177/0093650208330254.
[Crossref]

Zhao, Yuheng, Jinjing Jiang, Yi Chen, Richen Liu, Yalong Yang, Xiangyang Xue, and Siming Chen. 2022. Metaverse: Perspectives from Graphics, Interactions and Visualisation. *Visual Informatics* 6: 56–67. https://doi.org/10.1016/j.visinf.2022.03.002.
[Crossref]

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

Zuckerberg, Mark. 2021. *Meta.* https://about.facebook.com/meta/.

Part III
Against Algorithmic Democracy: Scope and Ethical-Discursive Perspectives for an Expansion of Deliberative Democracy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_8

8. Hyperethics: The Automation of Morality

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

If we do not want to eliminate the horizon of freedom and justice in contemporary societies, we must confront the technological determinism of such serious ethical-political consequences with an enlightened attitude (…). We must wake up from the new dogmatic dream with a progressive appearance and radically opt for freedom, without allowing ourselves to be seduced or subdued like herds by the current coercions of the new social power of a techno-political nature. We must abandon "optimistic techno-determinism", because digital technologies have created expectations that not only have not been fulfilled, but have worsened the situation for the development of authentic democracy.
Jesús Conill, Razón pública e inteligencia artificial (2022, 45)

Abstract
The aim of this chapter is to examine the defining characteristics of and challenges posed by the advent of a datafied, hyperconnected and algorithmic approach to the clarification, substantiation and application of morality: *hyperethics*. For this purpose, ethics will be addressed as practical knowledge concerned with the rationalization of free modes of behaviour, and which locates its criterion of morality within the dialogue between those concerned by an ethical issue. Attention will then be turned to an in-depth analysis of *ethification*, the current trend towards establishing processes for transforming social and moral reality into computable online data. It will then show how artificially intelligent mathematical models are progressively and relentlessly colonizing meaning-based rationalization processes, giving rise to meaninglessness, anomie and psychopathologies in mature democracies. Finally, we offer a critical reflection on the ethical and democratic challenges underlying the use of

Artificial intelligence as a tool to establish what is fair and pleasing for a digitally hyperconnected society.

Keywords Algorithmic colonization – Artificial and synthetic ethics – Lifeworld – Datafication – *Ethification* – Generative AI – *Synthetification*

8.1 Deliberation: The Dialogic Transformation of Morality

Throughout the history of philosophy, a distinction has been made between two types of science. On one hand there is formal science: a form of knowledge concerned with a general sense of understanding and the universal rules of thought, rather than empirical and experiential content; in other words, logic. On the other hand, material science, knowledge based on specific empirical and experiential content, has addressed the objects of the world and the universal laws that determine, condition or modulate their behaviour. In order to undertake this task, material science was divided into two subsets of knowledge: physics and ethics. The first of these, physics, concerns knowledge about the natural behaviour of objects and the universal laws regulating them. Meanwhile, the second, ethics, addresses the uninhibited behaviour of objects, those that escape the given order, the natural predetermination of things, and the universal laws that guide, regulate and modulate them. Thus, in Classical Greece, ethics were understood as knowledge about how human beings—the only objects of nature endowed with reason and action—make use of their freedom in relation to others, as well as the principles, values, precepts and virtues that they employ as a frame of reference to guide, regulate, modulate and criticize their own acts and decisions, as well as the acts and decisions of others.

Based on this idea, as set out in the earliest philosophical writings and enriched by contributions from other thinkers during a historical process lasting more than 2000 years, ethics can be defined as the practical knowledge that addresses the field of free relations and seeks to provide conceptual clarity about and establish the validity of the principles, precepts, actions and decisions at stake. It also applies the knowledge acquired in the different areas of activity in order to: resolve underlying conflicts, guide their development in a fair and pleasing sense, enhance their credibility and instil confidence in them, and increase their positive impact on society in order to minimize their vulnerability and make their plans for a good life possible (Conill 2006; Cortina 1986, 1990; Cortina and Martínez 1996; Cortina et al. 2003; García-Marzá 1992).

Among the various tasks undertaken by ethics, three stand out above the others. Firstly, there is the *conceptual clarification* of everything related to moral knowledge, such as the meaning and relationship between concepts like ethics, freedom, morality, democracy, the common good, civic virtue, professional excellence, intrinsic and extrinsic value, moral principle, moral precept, moral conscience, and so on. Secondly, the *foundation of morality* is the process of establishing the principles, values, precepts, duties, virtues and motives that make up the prescriptive framework of human relations at any given time and in any given place. And, thirdly, there is the application of all the baggage acquired in the previous two tasks so that what is just and pleasing is undertaken in the different spheres of human activity. This involves the introduction into the different spheres of activity to guide the appropriation, application and practical use human beings make of the principles, values, precepts, duties, virtues and motives for pursuing a good life based on freedom in relation to others.

Drawing on this definition and the distinctive tasks undertaken by ethics, different models of ethics coexist today, such as Aristotelianism, utilitarianism and deontology. Without underestimating the potential contributions that could be made by other approaches to ethics regarding the clarification, substantiation and application of moral knowledge, discursive approaches to deontological ethics such as that proposed by Karl Otto Apel and Jürgen Habermas in the 1980s (Apel 1980; Habermas 1987), and subsequently expanded and developed by Adela Cortina, Jesús Conill and Domingo García-Marzá (Cortina et al. 2003; Cortina 1986, 2007, 2010, 2022; Conill 2006, 2019; García-Marzá 1992, 2019) prove extremely fruitful as a means for mature democracies to successfully tackle the challenges raised by the moral development of a hyperdigital society across its diverse spheres of activity. In particular, through this renewed dialogic and cordial discursive ethics, it is possible to integrate the context and calculate the consequences of digitization. It also offers a means to address the reconstruction and promotion of both a culture or character that meets the standards required and desirable for a morally plural hyperdigital society, and the virtues and affections that channel, enable and promote it in practice.

Meanwhile, this discursive approach to ethics engages with the realm of intersubjectivity to offer a criterion of rationality—the intersubjective consensus of all those affected both at present and in the future concerning the acceptance of the consequences of a precept, action or decision that has been applied, or is applicable within a discourse given there are conditions of equality—for the purpose of critiquing both knowledge and action through dialogue. To this end, the proposal set out here is based on a process of dialogue intended to form an understanding about what is just, based on two fundamental principles: a

procedural moral principle (D) (within a practical discourse with specific rules and principles, the only precepts, actions or decisions that can be claimed as valid are those approved by all those affected); and a principle of universalization (U) (any precept, action or decision that claims to be valid must fulfil the condition that its effects, and that the foreseeable consequences can be accepted by all those affected). In turn, these two principles require a dialogue process to engage with an understanding of issues of justice based on three values: inclusion (all those affected at present and in the future should be able to participate in dialogues in which the precepts, actions or decisions that it will bring about or address are discussed); equality (all participants in these dialogues should be able to argue on equal terms in terms of their access to information, the absence of any form of pressure, and their being granted an equal time to speak, and so on); and reciprocity (all participants' interests should be considered equally and be open to critical review by others) (García-Marzá 1992).

On the other hand, this discursive approach to ethics has been extended and developed by introducing a cordial, hermeneutic, axiological and prudent resignation that enables the excessive proceduralism of the original proposal to be minimized. This means its applicability can be addressed across the different spheres of human activity (Cortina 1986; Conill 2006; García-Marzá 1992). In this sense, the proposal set out here addresses the moral validity of reciprocal expectations concerning forms of behaviour: the universally binding precepts that regulate the relational spaces occupied by communicatively linked free beings. In addition, the development of this proposal also opens up the possibility for the inclusion of concerns about the principles, values, virtues, aims, motives and affects that guide and motivate the behaviour of communicatively and affectively linked moral agents.

By deploying the focus provided by a renewed ethic of discourse, mature democracies could seek to discern the normative presuppositions underlying diverse fields of activity—the frame of reference for any attempt to justify them—in order to rationally criticize and guide both the aims and behaviour involved through arguments and dialogue. Likewise, this same focus will enable the establishment of methods for solving moral problems and conflicts, as well as establishing the civic virtues that underlie excellence of character within this sphere of action (Codina 2015). Furthermore, it provides a framework for developing the tools, mechanisms and guidelines that allow its practical application and fact-based re-creation.

However, at present this dialogic transformation of moral knowledge is being subjected to alteration; mathematical models are constantly intruding into the

meaning-based rationalization processes that mature societies use to reach intersubjective agreements about what is just or right, and more broadly to mitigate human vulnerability in the face of the given and ever-increasing complexity of the contexts they are related to. The results achieved within different fields by Artificial intelligence have generated a false belief that a *hyper-ethical*—data-driven, hyperconnected and algorithmized—approach to moral clarification, substantiation and enforcement is possible and desirable. The impacts of this have already begun to be noted by the most vulnerable groups within society. In other words, the transition from the dialogic transformation– communicative reason– to the digital transformation– algorithmic reason– of the moral sphere is being promoted in order to answer the three questions which, according to Kant, reason asks itself: what can be known, what ought to be done and what is to be expected (1968a, b [A 805, B 833).

8.2 Datafication: The Algorithmic Transformation of Morality

The growing spread of *datafication*, which is promoted by some large companies and government agencies due to the immense strategic and predictive potential it is supposed to offer (Mayer-Schöenberger and Cukier 2013), is one of the most noticeable effects of the application and synergistic convergence of IoT, Big Data and AI on the digitally hyperconnected society.

Datafication is a process that uses *cyber-physical ecosystems* and the huge amount of data and metadata they generate in order to deploy artificially intelligent mathematical models. These can track, segment and find behavioural patterns in a specific or general area, group or society to meet a strategic objective, such as a government's security, the education of its citizens, or the innovation and personalization of new products and services by an economic organization.

The success of *datafication* and its rapid assimilation into every sphere of human activity, has had much to do with the naïve, neutral or sugar-coated vision of it projected onto society by a range of governmental and economic organizations. Such discourses often highlight the social benefits to be obtained, such as greater security to prevent terrorist attacks, better control in the face of pandemics, substantial improvement in the diagnosis and cure of diseases, more efficient management of production processes, a considerable increase in the effectiveness of medicines, as well as the more sustainable use of scarce resources. It is even claimed that *datafication* gives rise to more objective,

inclusive and fairer decision-making. *Datafication* is seen by some sectors of society as "(…) a legitimate means to *access*, *understand* and *monitor* people's behavior" (van Dijck 2014), as the possible instrumental uses of its application on society are seen as a lesser evil compared to the great benefits it generates for society as a whole. It is, one might say, the latest and subtlest version of the metaphor of the *invisible hand*, *but now it is an algorithmic hand* that pulls the strings of new approaches to public policy, economics, democracy, medicine, education, among other spheres. Indeed, a stage has been reached at which an algorithmic reality has been established with its own reflexivity and legitimacy. We will now turn to consider a number of key examples.

Among the most important cases is the mass surveillance programme launched by the Chinese authorities to prevent, control or improve citizens' behaviour in a range of areas of social activity. This initiative is continually expanding; the *AI Global Surveillance Index* study, provides data on 176 countries that use this technology (Feldstein 2019). First and foremost, China's Social Credit System is particularly important: millions of cameras equipped with facial recognition technology placed in public and private spaces, such as streets, subways and rented houses, are used by an algorithm called *skynet* to collect behavioural information, measure the degree of civic conduct shown by the digitally *hyperconnected* citizens, and draw up a personalized good citizen card that allows or prevents access to public places, contracts or state aid (Laniuk 2021). Likewise, the "Intelligent Classroom Behaviour Management System" represents a striking departure: cameras equipped with facial recognition technology are placed on classroom blackboards and send information to an algorithm capable of identifying the students' degree of attention, as well as classifying their emotional state at any given moment (neutral, happy, sad, disappointed, annoyed, frightened or surprised). Teachers are alerted when the system identifies anomalies such as a lack of attention or signs of negative emotions (Brown et al. 2021).

Another significant development is the use of AI models by judicial and law enforcement agencies from around the world to decide on court sentences, enforce punishments and predict crimes, with regard to both offenders and victims (D'amato et al. 2020). Software such as *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) is being increasingly used; its algorithm performs statistical calculations based on offenders' online metadata and records of their behaviour to give advice to judges about whether a defendant is guilty, as well as the type and duration of the sentence to be imposed. It is also used to consider the likelihood of recidivism in the case of prisoners eligible for parole. A comparable

development is the use of predictive software, such as the *National Data Analytics Solution* (NDAS). Its algorithm performs statistical calculations based on the online metadata for the inhabitants of a particular neighbourhood or city, as well as any records of their behaviour, to enable police to make predictions about future crimes, offenders and victims.

Finally, attention must be paid to the use of algorithms equipped with AI that are used to create, develop, sell and distribute products and services. Amazon, for example, has developed an AI algorithm that uses its customers' online data and metadata, as well as their history of online searches and purchases, to predict what they will order in the future and send it to them before they even know they want to buy it. Similarly, companies such as Lantia Publishing have developed artificially intelligent mathematical models capable of monitoring and processing user behavioural data concerning the publishing and film industry to detect market trends and identify and develop successful products.

Academics, professionals and technologists concerned with the digital transformation of the moral sphere have begun to engage with the supposed achievements obtained by implementing processes for the *datafication* of social reality across the range of spheres of human activity, including politics, economics, education and health. This trend, known as *ethification or hyperethification*, which is based on *datafication* and underpinned by the supposed possibility of *hyperconnecting* social and human behaviour and *algorithmizing* the analysis, understanding and interpretation of phenomena (Calvo 2019a, 2021, Conill 2023), is a component of the *datafication* process. It endeavours to convert facts about the digitally *hyperconnected* citizenry, along with their opinions, preferences, habits and examples of their moral behaviour into online data and metadata. The aim is to develop a descriptive approach—what is—rather than a normative one—what should be—based on a strictly utilitarian criterion—the greatest good for the greatest number of people possible—in order to "(…) establish using artificially intelligent mathematical models what is morally desirable—*pleasing*—and valid—just—for society" (Calvo 2021). As Jesús Conill argues,

> The processes of "datafication" or "datification" are invading even the moral sphere, producing an "ethification" (…). Dialogue and critical reflection governed by claims of validity are relegated or annulled in favour of the aggregation of opinions, preferences and habits. Thus, communicative reason is disregarded or eliminated in favour of the calculation and mathematization of data, which serves to predict behaviour and feed instrumental and strategic reason. Reciprocal

> criticism between those who participate in the public sphere is thus eliminated, because it presupposes a conception of society oriented by quantitative individualism (Conill 2023, 124).

Regarding the automation of the maximums of happiness—moral—, the use of LLM as spiritual support to give guidance and meaning to the lives of citizens is particularly noteworthy. In this regard, in 2016 the Longquan Temple in Beijing (China) introduced Xian'er, a robot monk to spread wisdom and answer questions about Buddhism using its touchscreen (Sherwood 2016). In 2017, Germany introduced the BlessU-2 robot, which offers blessings in German, English, French, Spanish and Polish and answers questions from parishioners (Sherwood 2017). In 2022, Rabbi Joshua Franklin of the Hamptons Jewish Center acknowledged that he had used ChatGPT to generate sermons. And in 2023, ChatGPT delivered a mass in Germany, at St. Pauls Church in the city of Fuerth (Bavaria) during the "Deutscher Evangelischer Kirchentag" holiday (Bhuiyan 2023), and another in the USA, at Violet Crown Church in the city of Austin (Texas) (Murney 2023).

Regarding the automation of the minimums of justice (ethics), the principal engagements with *ethification* or *hyperethification* and its practical application have taken the form of various attempts to endow artificially intelligent mathematical models with moral competences and capacities by collecting massive volumes of data concerning the moral behaviour of citizens. Two landmark examples of this are the experiments undertaken by *The Moral Machine*, a platform developed by MIT's MediaLab to collect data on situations that pose moral dilemmas when driving a car; and *Moral Choice Machine*, a research project that seeks to equip machines with the necessary competences to "(…) learn to make ethical or even moral choices" by mining, quantifying and comparing cultural sources on moral choices (Jentzsch et al. 2019). A false belief provides the premise for these and similar projects and studies: it is possible for an AI-enabled machine to establish what is just and desirable in any given context by quantifying opinions, examples of behaviour and dilemmatic moral magnitudes.

Another manifestation of the emerging interest in *ethificiation* or *hyperethificiation* and its practical application are attempts to develop AI algorithms capable of making fair decisions based on the objectivity and neutrality of online data and metadata (Calvo 2019b, 2020a, b). This development is demonstrated by the algorithmic politicians already mentioned such as: Alisa—the Yandex Artificial intelligence system that attempted to run in the 2018 Russian elections claiming it was better than flesh-and-blood

politicians because it was not driven by emotions, did not seek personal advantage and did not make judgements on the interests and preferences of the citizenry; AI Mayor—the AI that ran for mayor of Tama City (Japan) in 2018, promising to end political corruption, engage in dialogue and reach an understanding with other political forces for the good of the district and provide "(…) fair and balanced opportunities for everyone" (Johnston 2018); and SAM —which sought to build support for a guaranteed 2020 New Zealand presidential election by offering the electorate an AI design capable of making fair, neutral and accountable decisions for the common good of citizens based on online data and metadata concerning the facts, behaviours and opinions of the electorate. As Tetsuzo Matsumoto, one of AI Mayor's backers in the 2018 elections, argues in *The Day AI Becomes God*: *The Singularity Will Save Humanity:*

> The following could perhaps describe an ideal implementation of democracy: it makes full use of Artificial intelligence to first determine public sentiment (dissatisfaction with the status quo and so on), identify the long-term advantages and disadvantages of the various available policy options, educate the population by explaining these options to them in a comprehensible manner, and first sounding out public sentiment before deciding on and implementing policies in accordance with the findings (Matsumoto 2018, 161).

A further manifestation of the emerging interest in *ethificiation* or *hyperethification* and its practical application is found in the field of health care, where efforts are underway to integrate artificially intelligent algorithms into processes involving moral deliberation. For example, ORACLE (Organizer of the Rational Approach in Computational Learning Ethics), an artificially intelligent mathematical model designed to assist bioethics committees in conflict resolution and ethical decision-making concerning clinical practice (Motta et al. 2016). The study "Investigation of the visual attention role in clinical bioethics decision-making using machine learning algorithms" (Fernandes et al. 2017) is another key example. It proposes using a machine learning algorithm and two data mining techniques to develop a predictive model that uses eye-tracking data to aid decision-making processes in clinical bioethics, especially those related to euthanasia.

Also noteworthy are cases on the design and use of artificial neural networks to improve the application of the principle of justice in the management of transplant lists (Attia, et al. 2023), the principle of non-maleficence in patients with advanced chronic disease (Ameneiros-Lago et al. 2015) or the principle of

autonomy through its implementation in the verbal and documentary processes of informed consent (Allen et al. 2023).

The current, increasingly widespread and unrestrained debate on the use of generative AI in moral decision-making and the management and resolution of ethical conflicts in the field of healthcare is particularly significant. A good example of this is the study "ChatGPT: aid to medical ethics decision-making?" In it, the authors reflect on the need and possibility of using LLMs as ChatGPT in ethical decision-making processes. Finally, they conclude that "Due to the rapid technical development and access to ever-increasing quantities of data, the utilisation should be closely observed and evaluated" (Schmidt and Lechner 2023). Also illustrative of the current context is the study "Consent-GPT: is it ethical to delegate procedural consent to conversational AI?" It reflects on the possibility and results of delegating the informed consent process from ChatGPT. In this regard, the study concludes "that at least in certain clinical situations, the benefits of LLMs potentially far outweigh those of current practices" (Allen et al. 2023, 1).

The foundation for this emerging interest in the *datafication* of social and moral reality is a scientific-technical ideology and a systemic economic-state imposition that results in major disruptions and anomalies affecting the proper development of a digitally *hyperconnected* society in the twenty-first century. The intrusion of Artificial intelligence algorithms into rationalization processes intended to give meaning to different things in the objective, social and subjective world, such as knowledge generation, human action, the construction of personality or the projection of a good life, is giving rise to a subtle but implacable colonization of the lifeworld through systemic imperatives currently based on scientific-technical concerns that are radically altering the structures and the proper functioning of modern societies across all their dimensions.

The most worrying impact of this development is the false belief that a model based on *datafication algorithmization* and *digital hyperconnectivity* capable of providing clarification, substantiation and application is both possible and desirable. Among other concerns, the resulting *hyper-ethical* approach could be used to legitimize social injustices, to reduce spaces for freedom and relations, misrepresent morally valid precepts, values, feelings, virtues and standards, in order to shift responsibility for actions and decisions onto *amoral agents* and displace any possible consensus among those affected through a computational analysis of the available subjective information. This is the hypothesis explored throughout the chapters of this block. Behind the apparent neutrality of *datafication* lies an ideology—a concealment of reality—whose clear aims are to achieve a form of technological domination unhindered by any

moral assessment from all those who have to suffer its consequences.

8.3 Hyperethics: The Algorithmic colonization of the Lifeworld

From a critical perspective, Habermas (1987) proposed that modern society is structured on the basis of two analytically differentiable but complementary and necessary levels: the functional level of the systemic world and the communicative level of the lifeworld. On one hand, the *systemic level* comprises two basic areas of coordination through the generalized media of the economy—money—and politics—power—whose projection and development depend upon a series of strategic actions, i.e. rationalization processes concerned with forging agreements on goals. On the other hand, the *lifeworld level* is made up of three spheres—culture, society and personality—which account for the objective world as the totality of entities about which it is possible to make true statements; the social world—as the totality of legitimately regulated interpersonal relations; and the subjective world—as the totality of the specific experiences lived by the subjects of society. The projection and development of these three spheres depend upon the implementation of communicative actions, i.e. rationalization processes mediated by an intersubjective understanding.

The lifeworld, which Habermas defines as "(…) culturally transmitted and linguistically organized stock of interpretative patterns" (1987, 124), represents the set of unquestioned or self-evident convictions that subjects endowed with speech and action use to enter into the *cooperative interpretation processes* whose primary aim is to form an understanding of different things in this world. In other words, from, and on the basis of, the lifeworld shared by speakers and listeners, both seek to understand one another with regard to something in the objective world—truth; in the social world—justice; and in the subjective world—veracity (1987).

In order to form an understanding, and thereby a definition of the social system as a whole (Habermas 1987), the system's distinct coordination mechanisms—money and power—must be anchored in the lifeworld to provide themselves with the meaning and legitimacy needed to justify their existence and permit them to endure (Habermas 1987). A government or a market that turns its back on the lifeworld is in danger of becoming blurred by deficits of meaning that threaten its survival and development through a lack of credibility and, therefore, trust. Nevertheless, as the lifeworld has become increasingly complex during the modern and contemporary eras as a result of the emergence of new

and greater demands for meaning, different spheres—culture, society and personality—have been colonized by systemic imperatives which, like power and money, produce anomalies (anomie, alienation and psychopathologies) that endanger its autonomy and thus its development and survival.

As Habermas argues, when violence (Gewalt) is introduced as an alternative to the mechanism for coordinating action, which is represented by understanding and power (Macht) as a product of action oriented towards understanding, it achieves the advantage of not losing sight of the forms of indirectly exercising violence that predominate today. Here, we are referring to the pathogenic violence that inadvertently seeps into the pores of everyday communicative practice and which can deploy its latent efficacy to the extent that the lifeworld becomes bound to the imperatives of functional subsystems autonomized and reified by the paths taken by unilateral rationalization (Habermas 1987).

Today, this has been exponentially aggravated by the processes of digital transformation. The new hyperconnected and hyper-datafied world that underlies these produces serious impacts on society. These include dependency on technology, imposition and alienation, disfigurement and dehumanization of human relations, techno-scientific transvaluation of democratic principles and values, reduction of social reality to a mere web of data sets and metadata, perception of human obsolescence and algorithmic neutrality, techno-chauvinism and optimistic techno-determinism, intellectual and cognitive regression, and the algorithmization of inequality and injustice.

The latter development has led to the systemic colonization of the lifeworld becoming determined by an ever-greater technification and instrumentalization of its structural components through the *datafication of* symbolic production across all its dimensions. The result of this *hyper-datafied* world is the imposition of "(…) a perspective in which the development of the social system seems to be determined by the logic of scientific-technical progress" (Habermas 1987, 279–280), and the reduction of linked decision-making processes to the mere calculation of the series of available data and metadata.

Following Habermas, the main problem faced by a world increasingly governed by Artificial intelligence algorithms, and thereby more dependent upon *datafication* and *digital hyperconnectivity* processes, is that calculation taken to the extreme—the principal characteristic of the decision-making processes engaged in by this type of scientific-technical perspective—decisions in a state of purity as regards the cleanliness of all elements that could still be considered accessible to any binding analysis (Habermas 1987). As a result, *datafication* creates an uncritical and anomically deficient evolutionary process within the social system that gives rise to serious interferences in symbolic reproduction,

meaning generation and the coordination of action.

A clear indicator of this is the incipient interest shown by a growing and influential number of academics, technologists and sympathizers in concretizing and applying an *algorithmized* model of clarification, substantiation and/or application of morality: *hyperethics*. This will lead to serious anomalies in rationalization processes concerning the meaning of the social world. *Ethification* or *hyperethification*—understood as a digitized method for compiling and generating moral knowledge in a *broad* and *objective* way through the use of ICTs, as well as implementing *datafication* processes in the moral sphere—is intended to identify in real time the actions, precepts and decisions that are morally just and felicitous for society, the virtues that inform excellent citizenship, the feelings and interests that constitute motivations for moral action in societies, and the competences and capacities that allow their fact-based re-creation, among other things.

The prefix *hyper*, as in economics, is used in a triple sense. Firstly, in the sense of *hypertextuality*, as a structured set of *things*, as well as people, bound together by links and connections. Then, in the sense of *hyper-information* and *hyperhistory*, as social facts and processes that are highly dependent on ICTs and their results. Finally, in terms of *hyperspace*, as a representation of a *virtualized public sphere* that apparently escapes the logic of deliberation and the search for agreements.

In this regard, the specialized literature contains numerous studies and experiments (Allen and Wallach 2009, 2011; Awad et al. 2018; Bostrom 2014; Floridi and Sanders 2004; Hooker 2018; Jentzsch et al. 2019; Matsumoto 2018; Motta et al. 2016; Siqueira-Batista et al. 2014; Tigard 2021) that consciously or unconsciously contribute to the promotion and development of *hyperethics*. They are based on a number of basic concepts and characteristics:

1. *Artificial moral meaning:* As a starting point, *hyperethics* replaces deliberative processes that lead to those affected by an applied or applicable precept, action or decision trying to understand the moral validity of something in the world through dialogue and possible intersubjective agreement. Instead, it offers processes based on artificially intelligent mathematical models which undertake the computational analysis of huge databases and metadata. AI captures the *moral sense* of things in the world by computing all available data and metadata. Using this method, AI extracts recurring sets of moral items, attributes or magnitudes; in other words, moral patterns, which it then uses to guide society concerning what

is just and pleasing, or else for the AI to use as a criterion of rationality in the decision-making processes it participates in. Moral meaning is therefore reduced to a description and classification of behavioural reality in society: an aggregate of the displays of willingness, behaviour and opinions encountered on line.

2. *Artificial moral agency: Hyperethics* extends the boundaries of moral agency to include certain amoral entities, such as mathematical models endowed with AI. For example, in their study "On the Morality of Artificial Agents" (2004), Luciano Floridi and J.W. Sanders argue for the morality of AI-enabled algorithms that make decisions. For this purpose, they apply a concept of moral agency—one which sidesteps issues such as freedom, mental states and the agent's responsibility—accompanied by a discourse that subjugates ethics to law and reduces freedom to a mere process of free will, among many other factors. The development of these types of perspective has fostered the current use of AI algorithms in decision-making processes across different spheres of social activity for the purpose of circumventing accountability.

3. *Artificial moral conditionality*: *Hyperethics* remains totally dependent on ICTs and the huge flow of data and metadata. Were these ICTs to malfunction or be disconnected, in the result would be the paralysis of this approach due to a lack of, or scarcity of, data for processing to be converted into relevant information and applicable knowledge. The results would thus be distorted or arbitrary. The partial or general *moral connection* or *disconnection* of this type of approach is therefore conditioned and/or influenced by the strategic decisions taken by Big Tech companies, as it is they who control the flow of online data and metadata, oversee most of the major databases and grant access to the different levels of hyperconnectivity, allowing them to influence the AI algorithms involved to a greater or lesser extent.

4. *Artificial moral objectivity*: *Hyperethics* formulates a descriptive proposal for the clarification, substantiation and application of morality. This *datafied*, hyperconnected and algorithmic approach to ethics does not seek to prescribe social reality in order to channel and improve it, but to describe the facts and opinions, customs and behaviours of digitally *hyperconnected* citizens in order to obtain statistical patterns that enable it to decide which decisions, patterns and modes of behaviour are *objectively* just and should

be congratulated.

5. *Artificial moral reality: Hyperethics* promotes a binary structure of reality for resolving conflicts, coordinating action and making decisions. Based on moral dilemmas and focused on the outcomes of actions and decisions, the proposal reduces all reality to just two possibilities in order to overcome the huge complexity involved. In this way, as can be seen in experiments such as *The Moral Machine* (2018), when faced with a specific event, this approach offers two exclusive possibilities (the choice of one requires the other to be forgotten) and a *rational* solution (based on the available data and metadata and the calculation of foreseeable consequences, which, in most cases, tend to be immoral).

As a result, *hyperethics* converts prescriptive and emancipatory moral knowledge into a predictive and deterministic scientific-technical knowledge. This epistemic conversion interferes in the specification of morally valid and pleasing precepts, values, feelings, virtues and guidelines, which gives rise to deficits in legitimacy, a scarcity of binding precepts, axiological poverty, a decline in virtue, and disaffection with participation, all of which endangers the rationality, solidarity and autonomy of the lifeworld (Habermas 1987). Together, this makes *hyperethics* an apt mechanism for the legitimization of social injustice and the persistence of morally undesirable patterns and behaviour for societies with a post-conventional level of moral development that limits progress across their different spheres. As Habermas noted, it is the immanent legitimacy of such processes that seems to produce the specific material coercions to which a policy aimed at meeting functional needs must conform. And when this appearance has been effectively imposed, then resorting propagandistically to the role of science and technology can explain and legitimize why in modern societies a democratic formation of political will in relation to practical questions has lost its functions, and can be replaced by plebiscitary decisions concerning alternative teams of administrators (Habermas 1987).

Thus, there is a close and inalienable link between moral knowledge and democracy. The *hyper-ethical* approach to the clarification, substantiation and application of morality attacks the meaning-based rationalization processes that legitimize democratic processes and their impact, It therefore represents a major challenge that needs to be seriously and rigorously addressed.

8.4 The Ethical and Democratic Challenges Posed by *Hyperethics*

According to an ethical perspective framed by the universalist, deontological, proceduralist, minimal, hermeneutic, dialogic and critical approach proposed by Karl Otto Apel and Jürgen Habermas in the 1980s, (Apel 1980; Habermas 1984a, 1987), whose scope has been developed and expanded, above all, by Adela Cortina, Jesús Conill and Domingo García-Marzá (Cortina 1986, 1990, 2007, 2010, 2022; Cortina 2007; Conill 2006, 2019; García-Marzá 1992, 2016), moral knowledge cannot be reduced to the aggregation and computational processing of online data and metadata on the interests, opinions and behaviour of a digitally hyperconnected society. Above all, this is because this would be to disregard citizens' autonomy, as well as their ability to criticize a state of affairs, deliberate upon the best courses of action for conflict resolution, reach agreements on how to coordinate action, convince others with sound reasons concerning the *ligatio* that *ob-liges* everyone to act in a certain way, and make changes in order to improve, revise or eradicate different things and forms of behaviour, even when this goes against their own interest. In other words,

> (…) an autonomy expressed in our everyday moral language, which allows us to talk about mistaken interests and to undertake an argued critique of the decisions taken by our representatives, even though they may form the majority. The majority does not imply truth, let alone justice. This becomes quite clear when we refer to the injustice of certain aggregations that leave "out" a part of the citizens, or when we think about the inheritance we are leaving to future generations (García-Marzá 2015: 102).

In this sense, "a theory of democracy cannot ignore this moral knowledge, these capacities or resources, this ethical capital" (García-Marzá 2015: 102), as this is a condition for its legitimacy and credibility, and therefore its survival and development. Indeed, this is precisely the source of the deliberative democracies which we will go on to argue are conducive to and productive for addressing the challenges of the new digital contexts: public deliberation by free and equal citizens is the key to the legitimacy of the political process, as these citizens would acknowledge the existence of a moral core at the basis of all possible legitimation and engage in democracy as a power to transform people's interests in a morally acceptable way. And this power cannot be achieved without the free

and equal participation of all those affected (García-Marzá 2015): 102).

This nucleus is underpinned by participation as a moral requirement derived from autonomy. From a moral point of view, the real and effective participation of all those affected is required, as there is no other way to speak of justice other than through the free and equal agreement of all those involved. As soon as a political process does not consider preferences and interests as something given, but as capable of being discursively transformed, responsibility can never be delegated; in other words, there is no possibility of participating in order to address one's own interests. The participation required must therefore be real, direct and effective.

Therefore, from an ethical perspective such as the deliberative and cordial one upheld here, the uncritical acceptance and application of a *datafied*, algorithmic and *hyperconnected* approach to the clarification, substantiation and application of morality can generate unintended or intentional consequences for society and citizens, and likewise for the mature democracies that base their emergence and development on the participation of those affected by the impact of this approach. Seven consequences are particularly important: reification, inequality, heteronomy, conventionalism, instrumentalization, regression and perversion.

1. Reification: *Hyperethics* is an approach to ethics that reduces the human being to a mere terminal contributing to the massive flow of online data and metadata. For *hyperethics*, the human being is seemingly nothing other than a *hyperconnected thing* within a *cyber-physical ecosystem*, and their value is no longer seen as intrinsic and linked to its ability to self-legislate, but extrinsic, and linked to its ability to be hyperconnected and generate large volumes of data and metadata online. *Hyperethics* is therefore shown to be a fallacious approach that presents human beings as products, whose value is conditional on the amount of data they contribute to the system, rather than as ends in themselves, whose dignity is unconditional.
2. Inequality: The focus provided by *hyperethics* increases social complexity and, consequently, the division arising from all the dimensions of social inequality. *Hyperethics* is dependent on ICTs and the major inequalities undermining citizens' access to and use of such technologies. It also contributes a misogynist, aporophobic, homophobic and xenophobic bias to the mass volumes of online data it uses for its purposes, making this ethical model a malleable mechanism conducive to generating biased or manipulated descriptions of moral knowledge and, in so doing, providing

solutions of a sort that benefit the privileged and punish the most vulnerable in society.

3. Heteronomy: *Hyperethics* relies on artificial processes of discernment and decision-making on moral issues resulting in an alarming rise in disaffection with regard to participation and disinterest shown towards injustice. The false idea that AI algorithms are better equipped than people to discern what is valid and desirable in a particular or general society is taking root in contemporary society, and as a result a dangerous trend towards the adoption of paternalistic patterns of behaviour is emerging, within which AI controls all the decision-making power and humans uncritically abide by its dictates. This trend encourages a retreat from the socio-political project of modernity, within which the intellectual coming of age of the human being is the cornerstone of its emergence and development. Within the new socio-political project that seemingly emanates from digitally *hyperconnected* societies, the algorithmic governance of rationalization processes with respect to ends and, more worryingly, with respect to meaning seems to be a looming reality.

4. Conventionalism: *Hyperethics* is an approach whose conventionalism hampers the possibilities of transforming social reality in a just and pleasing sense. Mathematical models endowed with AI lack the necessary skills to critique what they encounter and to differentiate between the applicability and moral validity of a precept, action or decision. On one hand, the incompetence of AI algorithms prevents the establishment and projection of a better possible world for a particular or general society. Meanwhile, on the other hand, this algorithmic incompetence encourages the continuity of anachronistic patterns and behaviours that are unacceptable for a mature society with a post-conventional level of moral development.

5. Instrumentalization: Hyperethics is an approach that promotes highly opaque processes and outcomes that violate the principle of openness. While ethics demands intelligibility, transparency and explicability, hyperethics hides its results behind computer language, massive data and metadata, and property rights, among other things. As discussed above, if ethics are not public, it is not for practical reasons, but for instrumental and strategic reasons. In other words, if the reasons for ethics cannot be public for whatever reason, it is most likely an attempt to conceal partisan ends.

6.

Regression: Hyperethics is an approach that encourages processes of regression towards the intellectual minority of citizens who violate the principle of autonomy. While ethics are based on people's ability to self-legislate, to dictate and enforce their own rules in relation to others, *ethification* is based on the computational superiority of the machine to process Big Data and metadata in order to apply authoritarian, protectionist and *heteronomising* practices concerning how people should fairly and felicitously guide their behaviour. In this way, "the exercise of reason and freedom is endangered, threatening the foundations of the enlightened conception of man and society, as well as constitutional democracy" (Conill 2022, 44).

7. Perversion: *Hyperethics* is an approach that perverts the processes of constructing moral knowledge and the occurrence of what is just. What AI algorithms do or can achieve is not ethics, but mere conventional, analytical, formal and descriptive information on the state of the question of the moral dimension. They rely on the Big Data and metadata that citizens produce about morality to produce behavioural patterns about what is just and worthy of congratulation for society. However, ethics are not based on data and metadata, conventional information and behavioural patterns. They are prescriptively and gradually constructed through post-conventional arguments, critical-discursive processes and intersubjective agreements where all those affected by an action, norm or decision accept the actual or virtual consequences of its practical application.
8. *Synthetification*: *Hyperethics* is an approach whose contents are constructed through processes of synthesising reality and moral justification (*synthetification*). Algorithms, especially those endowed with Generative AI, are colonising the digital public sphere and flooding it with synthetic datasets about what is just and congratulatory (synthetic opinion). This is distorting the processes of public opinion generation, transforming morality into synthetics and the public sphere into a kind of *algosphere*; that is, into a space where citizen participation and influence diminishes as the participation and influence of algorithms becomes massive, hegemonic and despotic.

These and other issues pose a challenge for the digitally hyperconnected society and thus for mature, pluralistic democracies with a post-conventional level of moral development. The digital transformation of society seems

irreversible, and it can provide major benefits. However, its emergence and development require a constant critique and the establishment of the appropriate guidelines to prevent an increase in social complexity and, as a result, the vulnerability of people, especially those belonging to the most disadvantaged groups in society. To achieve this, it is necessary to guide digital practice in a fair and responsible direction through the design and application of suitable frames of reference, such as guidelines and codes of ethics, conduct and good practice; the establishment of mechanisms for deliberation and dialogue concerning those affected by its consequences, including ethics committees; the implementation of instruments to capture information on the fulfilment of the commitments reached —for example, ethical approaches; and the implementation of accountability tools, such as explicability reports (García-Marzá 2017).

In short, one of the most worrying results of this new hyperdigitized world is the current trend towards *hyperethics* (artificial and synthetic ethics). This *data-driven* approach to the clarification, substantiation and application of morality is gradually attracting the interest of the scientific-technical and political spheres due to its ability to redress the oversight in the argument resulting from the increased complexity of the new hyperdigitized world. Among other things, the development of this type of proposal is particularly alarming due to its capacity to legitimize social injustices, to distort morally valid precepts, values, feelings, virtues and guidelines; to objectify and instrumentalize the citizenry; and to conceal or deflect responsibility for actions and decisions.

The ethical approach underlying the datafication of the moral sphere is not only fallacious, as a result of its basis on pragmatic and undecidable contradictions, it also undermines the rational processes of process of making sense of things that people establish in order to understand themselves about what is right in the world (Habermas 1987). The main reason for this is that *hyperethics* promotes the replacement of deliberation and any possible consensus among those concerned with a computational analysis of the available subjective information. Today, the shortcomings resulting from the increased complexity and new demands placed on communication within the hyperdigital world are being filled by Big Data, not dialogue. In conjunction with this, the life world (Lebenswelt) is being progressively colonized by AI through systemic imperatives, such as the power and money deployed by the most powerful states and Big Tech.

As a result, *artificial and synthetic hyperethics* turns prescriptive and emancipatory moral knowledge into predictive and deterministic scientific-technical knowledge, leading to interference in the specification of morally valid and *congratulatory* precepts, values, feelings, virtues and guidelines. This gives

rise to deficits in legitimacy, a scarcity of binding precepts, axiological poverty, a decline in virtue and a participatory disaffection that endangers the rationality, solidarity and autonomy of the lifeworld (Habermas 1987). It is these combined factors that make *artificial and synthetic hyperethics* a suitable mechanism for legitimizing social injustices and sustaining morally undesirable models and forms of behaviours in societies with a post-conventional level of moral development, and this limits any progress across their different spheres.

Bibliography

Akiyama, Kazunori, et al. 2019. First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole. *The Astrophysical Journal Letters* 875 (L4): 1–52.

Allen, C., and W. Wallach. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

———. 2011. Moral Machines: Contradiction in Terms or Abdication of Human Responsibility? In *Robot Ethics: The Ethical and Social Implications of Robotics*, ed. P. Lin, K. Abney, and G.A. Bekey, 55–68. Cambridge, MA: MIT Press.

Allen, J.W., B.D. Earp, J. Koplin, and D. Wilkinson. 2023. Consent-GPT: Is It Ethical to Delegate Procedural Consent to Conversational AI? *Journal of Medical Ethics* 0: 1–7. https://doi.org/10.1136/jme-2023-109347. Published Online First.
[Crossref]

Ameneiros-Lago, E., C. Carballada-Rico, J.A. Garrido-Sanjuán, and A. García Martínez. 2015. Adecuación de las intervenciones clínicas en pacientes con enfermedad avanzada y compleja. Propuesta de algoritmo para la toma de decisiones. *Revista de Calidad Asistencial* 30 (1): 38–43.

Apel, K.O. 1980. *Towards a Transformation of Philosophy*. London & Boston: Routledge & Kegan Paul.

Attia, Antony, Ian A. Rowe, Ewen M. Harrison, Tim Gordon-Walker, and Ben M. Stutchfield. 2023. Implausible Algorithm Output in UK Liver Transplantation Allocation Scheme: Importance of Transparency. *The Lancet* 401: 911–912.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine Experiment. *Nature* 563: 59–64.

Bhuiyan, Johana. 2023. Are Chatbots Changing the Face of Religion? Three Faith Leaders on Grappling with AI. *The Guardian*. https://www.theguardian.com/technology/2023/apr/07/chatgpt-artificial-intelligence-religion-faith-leaders.

Boden, Margaret A. 2017. *Inteligencia Artificial*. Madrid: Turner Noema.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brown, T.G., A. Statman, and C. Sui. 2021. Public Debate on Facial Recognition Technologies in China. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, https://doi.org/10.21428/2c646de5.

37712c5c.

Calvo, Patrici. 2019a. Etificación, la transformación digital de lo moral. *Kriterion. Revista de Filosofía* 60 (144): 671–688.

———. 2019b. Democracia algorítmica: consideraciones éticas sobre la *dataficación* de la esfera pública. *Revista del CLAD. Reforma y Democracia* (74), 5–30.

———. 2020a. Gobierno algorítmico: sobre el neuroaprendizaje moral de las máquinas en la política y la economía. In *Neuroeducación moral y Democracia*, ed. Jesús Conill and Domingo García-Marzá, 223–242. Comares: Granada.

———. 2020b. Democracia aumentada: un *ecosistema ciberético* para la participación política basada en algoritmos. *Apeiron. Revista de Filosofía* (12), 129–141.

———. 2021. Etificiación. In *Global Observatory for Social Pathologies*, ed. Jovino Pizzi and Maximiliano Cenci. Pelotas: Federal University of Pelotas.

Center for the Governance of Change. 2019. European Tech Insights 2019. Mapping European Attitudes to Technological Change and its Governance. http://docs.ie.edu/cgc/European-Tech-Insights-2019.pdf.

———. 2020. European Tech Insights 2019. Mapping European Attitudes to Technological Change and Its Governance. http://docs.ie.edu/cgc/European-Tech-Insights-2019.pdf.

———. 2021. *European Tech Insights 2019. Mapping European Attitudes to Technological Change and its Governance*. http://docs.ie.edu/cgc/European-Tech-Insights-2019.pdf.

Codina, María José. 2015. *Neuroeducación en virtudes cordiales: cómo reconciliar lo que decimos con lo que hacemos*. Barcelona: Editorial Octaedro.

Conill, Jesús. 2006. *Ética Hermenéutica. Crítica desde la Facticidad*. Madrid: Tecnos.

———. 2019. *Intimidad corporal y persona humana. De Nietzsche a Ortega y Zubiri*. Madrid: Tecnos.

———. 2022. Razón pública e Inteligencia artificial. In *Actualidad de John Rawls en el siglo XXI*, ed. G. Pereira, y P.J. Pérez Zafrilla (eds.), 37–48. Granada: Comares.

———. 2023. Ética discursiva e Inteligencia artificial. ¿Favorece la inteligencia artificial la razón pública? *Daimon. Revista Internacional de Filosofía* 90: 115–130.

Cortina, Adela. 1986. *Ética mínima*. Madrid: Tecnos.

———. 1990. *Ética sin moral*. Madrid: Tecnos.

———. 2007. *Ética de la razón cordial. Educar en la ciudadanía en el siglo XXI*. Oviedo: Nobel.

———. 2010. *Justicia Cordial*. Madrid: Trotta.

———. 2022. *Aporophobia: Why We Reject the Poor Instead of Helping Them*. Cambridge: MIT Press.

Cortina, Adela, y Martínez, Emilio. 1996. *Ética*. Madrid: Akal.

Cortina, Adela, Domingo García-Marzá, and Jesús Conill. 2003. *Public Reason and Applied Ethics. The*

Ways of Practical reason in a Pluralist Society. New York: Routledge.

D'amato, C., I. Silver, J. Newsome, and E. Latessa. 2020. Progressing Policy Toward a Risk/Need Informed Sanctioning Model. *Criminology and Public Policy 20*: 41–69.

Dasgupta, Subrata. 2018. *Second Age of Computer Science*. Oxford: Oxford University Press.

de Mántaras, Ramon López, and Pedro Meseguer González. 2017. *Inteligencia Artificial*. Madrid: Cataract.

de Souza Motta, Luís Claudio, Lucas Nicolau de Oliveira, Eugenio Silva, and Rodrigo Siqueira-Batista. 2016. Toma de decisiones en (bio)ética clínica: enfoques contemporáneos. *Revista de Bioética* 24 (2): 304–331.

Dewitt, Barry, Baruch Fischhoff, and Nils-Eric Sahlin. 2019. Moral Machine' Experiment Is No Basis for Policymaking. *Nature* 567: 31.

Donati, Pierpaolo. 2019. *Sociología relacional de lo humano*. Barañain: EUNSA.

du Sautoy, Marcus. 2019. *The Creativity Code. How Ai is Learning to Write, Paint and Think*. London: 4th State.

Feldstein, Steven. 2019. *The Global Expansion of AI Surveillance*. Washington, DC: Carnegie.

Fernandes, Daniel L., Rodrigo Siqueira-Batista, Andréia P. Gomes, Camila R. Souza, Israel T. da Costa, Felippe da S.L. Cardoso, João V. de Assis, Gustavo H.L. Caetano, and Fabio R. Cerqueira. 2017. Investigation of the visual attention role in clinical bioethics decision-making using machine learning algorithms. *Procedia Computer Science* 108: 1165–1174.

Floridi, Luciano. 2012. Hyperhistory and the Philosophy of Information Policies. *Philosophy & Technology* 25: 129–131. https://doi.org/10.1007/s13347-012-0077-4.
[Crossref]

Floridi, Luciano, and J.W. Sanders. 2004. On the Morality of Artificial Agents. *Minds & Machines* 14: 349–379.

García-Marzá, Domingo. 1992. *Ética de la justicia: J. Habermas y la ética discursiva*. Madrid: Tecnos.

———. 2015. The Democratic Value of Civil Society: A Response to Disaffection", *THÉMATA. Revista de Filosofía (*52): 93–109. https://doi.org/10.12795/themata.2015.i52.05.

———. 2016. Applied Neuroethics: The Practical Consequences of Neuropositivism. *Pensamiento. Journal of Research and Technological Innovation* 72 (273): 881–900.

———. 2017. From Ethical Codes to Ethical Auditing: An Ethical Infrastructure for Social Responsibility Communication. *The Information Professional* 26 (2): 268–276.

———. 2019. Ética y democracia. Notas para una renovada ética del discurso. In *Ética y democracia desde la razón cordial*, eds. E. González, J. C. Siurana, J.L. López-González, y M. García-Granero (coords.), 7–16. Granada: Comares.

Habermas, Jürgen. 1984a. *Theory of Communicative Action Vol. 1: Reason and the Rationalization of Society*. Boston: Beacon Press.

———. 1984b. *Ciencia y técnica como "ideología"*. Madrid: Tecnos.

———. 1987. *Theory of Communicative Action Vol. 2 Lifeworld and System: A Critique of Functionalist Reason*. Boston: Beacon Press.

Hooker, John. 2018. Truly Autonomous Machines Are Ethical. *arXiv:1812.02217 [cs.AI]* https://arxiv.org/pdf/1812.02217.pdf.

Jentzsch, S., P. Schramowski, C.A. Rothkopf, and K. Kersting. (2019, 27–28 January). Semantics Derived Automatically from Language Corpora Contain Human-Like Moral Choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 37–44.

Johnston, Lachlan. 2018. There's an AI Running for the Mayoral Role of Tama City, Tokyo. *O taquest*, http://www.otaquest.com/tama-city-ai-mayor/.

Kant, Immanuel. 1968a. *Kritik der reinen Vernunft* (1. Auflage 1781), Kants Werke. Berlin: Akademie Textausgabe, Walter de Gruyter, IV, 1–252.

———. 1968b. *Kritik der reinen Vernunft* (2. Auflage 1787), Kants Werke. Berlin: Akademie Textausgabe, Walter de Gruyter, IV, 1–252.

Laniuk, Yevhen. 2021. Social Credit System as a Panopticon: Surveillance and Power in the Digital Age. In *Community and Tradition in Global Times*, ed. Denys Kiryukhin, 211–234. Washington, DC: Cultural Heritage and Contemporary Change.

Marco, María Jesús, Josep María Marco, Josep Prieto, and Ramón Segret, eds. 2010. *Escaneando la informática*. Barcelona: UOC.

Matsumoto, Tetsuzo. 2018. *The Day AI Becomes God. The Singularity will Save Humanity*. Cambridge (NZ): Media Tectonics.

Mayer-Schöenberger, Viktor, and Kenneth N. Cukier. 2013. *Big Data. A Revolution that Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

Motta, Luís Claudio de Souza, Lucas Nicolau de Oliveira, Eugenio Silva, and Rodrigo Siqueira-Batista. 2016. Toma de decisiones en (bio)ética clínica: enfoques contemporáneos. *Revista de Bioética* 24 (2): 304–331.

Murney, Michael. 2023. Texas Church Uses ChatGPT. *Chron*. https://www.msn.com/en-us/news/other/texas-church-uses-chatgpt-to-generate-material-for-sunday-service/ar-AA1h0T8W.

Norrie, T., et al. 2021. The Design Process for Google's Training Chips: TPUv2 and TPUv3. *IEEE Micro* 41 (2): 56–63.

PwC. 2021. Global Digital Trust Insights Survey 2020. https://www.pwc.es/es/publicaciones/transformacion-digital/global-digital-trust-insights.html.

Quan-Haase, Anabel, and Barry Wellman (2005, 16–20 February), How Computer Mediated Hyperconnectivity and Local Virtuality Foster Social Networks of Information and Coordination in a Community of Practice, paper presented at the *XXV International Sunbelt Social Network Conference*, Redondo Beach, California.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning Representations by Back-Propagating Errors. *Nature* 323: 533–536.

Schmidt, K.W., and F. Lechner. 2023. ChatGPT: Hilfe bei der medizinethischen Entscheidungsfindung? [ChatGPT: Aid to Medical Ethics Decision-Making?]. *Innere Medizin (Heidelb)* 64 (11): 1065–1071.

Schramowski, Patrick, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. 2019. BERT has a Moral Compass: Improvements of ethical and moral values of machines. *arXiv:1912.05238v1* https://arxiv.org/pdf/1912.05238.pdf.

Sherwood, Harriet. 2016. Robot Monk to Spread BUDDHIST WISDOM to the Digital Generation. *The Guardian*. https://www.theguardian.com/world/2016/apr/26/robot-monk-to-spread-buddhist-wisdom-to-the-digital-generation.

———. 2017. Robot Priest Unveiled in Germany to Mark 500 Years Since Reformation. *The Guardian*. https://www.theguardian.com/technology/2017/may/30/robot-priest-blessu-2-germany-reformation-exhibition.

Siqueira-Batista, Rodrigo, Andréia Patrícia Gomes, Polyana Mendes Maia, Israel Teoldo da Costa, Alcione Oliveira de Paiva, and Fábio Ribeiro Cerqueira. 2014. Los modelos de toma de decisiones en bioética clínica: apuntes para un enfoque computacional. *Revista de Bioética* 22 (3): 456–461.

Tigard, D. 2021. Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Cambridge Quarterly of Healthcare Ethics* 30 (3): 435–447.

van Dijck, José. 2014. Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology. *Surveillance and Society* 12 (2): 197–208.

Wallach, Wendell, Colin Allen, and Stan Franklin. 2011. Consciousness and Ethics: Artificially Conscious Moral Agents. *International Journal of Machine Consciousness* 3 (1): 177–192.

Writer, Beta. 2019. *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research*. Cham: Springer.

Zhang, Xian-Da. 2020. Machine Learning. In *A Matrix Algebra Approach to Artificial Intelligence*, ed. Xian-Da Zhang. Singapore: Springer.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_9

9. A Critique of Algorithmic Reason

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied an entire City, and the map of the Empire, an entire Province. In time, these Excessive Maps did not satisfy and the Schools of Cartographers built a Map of the Empire, that was of the Size of the Empire, and which coincided point for point with it. Less Addicted to the Study of Cartography, the Following Generations understood that that dilated Map was Useless and not without Pitilessness they delivered it to the Inclemencies of the Sun and the Winters. In the Deserts of the West endure broken Ruins of the Map, inhabited by Animals and Beggars; in the whole country there is no other relic of the Disciplines of Geography.
Jorge Luis Borges, Del rigor de la ciencia (1960)

Abstract
The pervasive force with which the current *algorithmic colonization* has spread across both the state and civil society is the result of an attitude and state of mind, a sense of confidence and naivety that can only be explained by referring to a prior cultural context capable of defining and paving the way for what is now underway. The aim of this chapter is to analyse the cultural and political foundations that have enabled the rapid spread of AI across all spheres of life, in other words, what has come to be known as the algorithmic revolution. To a considerable extent mistrust and inequality explain the current democratic drift, but algorithmic reason has engaged with this situation through its theoretical scientific approach and technological application to the different spheres of the state and civil society. However, these two facets of algorithmic development require and presuppose the application of evaluative criteria of legitimacy, or to

express it in simpler terms, consistency and a certificate of validity in order to operate. The ongoing accelerated phase of algorithmic colonization is impossible to understand, and still harder to explain, unless we can comprehend what preceded it: a terrain fertilized by scientism or objectivism. However we may wish to refer to it, for years this terrain has been reducing human development to technological development, and now, as if amid a perfect storm, it has now been supplemented with neuropower and post-truth.

Keywords Algorithmic culture – Algorithmic neutrality – Algorithmic ideology – Datapositivism – Neuropower – Post-truth – Platforms

---

9.1 Datapositivism: The New Positivism

The digital revolution wields considerable symbolic and technological power, which allows it to construct meaning and credibility for its activities. It takes considerable effort to relate two situations that have apparently occurred by coincidence and established the firm foundations upon which the aspirations of algorithmic democracy are currently based.

On one hand, as has been said, we have democratic disaffection, a scenario in which democratic systems suffer from widespread disaffection, distrust, and even indignation and deep resentment. We have discussed the causes of this phenomenon, which are almost always related to a failure to fulfil legitimate expectations of freedom and aspirations to equality. The current systems of liberal democracy and social commitment have proved incapable of coping with the two recent crises, the financial collapse and then the pandemic, and today, it is deeply unfortunate that they are also unable to cope with war in Europe itself: Russia's invasion of Ukraine.

On the other hand, contemporary society displays a firm, if not unshakeable, confidence in new technologies, in datafication and its basic elements such as Big Data, algorithms and the Internet of Things, and their capacity to solve the problems that it is claimed people's innate selfish individualism, interests and passions, have provoked. Algorithms will not only save us from our politicians, they will also save us from ourselves. Our current view of democracy, both in politics and in businesses, universities, hospitals, and so on, is being structured by an unjustified technological optimism.

According to Edmund Husserl's *Crisis of European Sciences and Transcendental Phenomenology*, enthusiasm for technological progress—in this case datafication—goes back at least as far as Galileo. When Husserl delivered the famous lecture that closes the book, he was already being persecuted by the

Nazis. And it was from this perspective of war—unfortunately very close to us today—that he questioned the significance of science for mankind in the past and present. Confronted by the "spiritual hatred and barbarism" then pervading Europe, Husserl confirmed that the sciences had fallen silent; they had lost all contact with human reality, having become imprisoned by a methodology that sought to imitate the success of the empirical sciences. Husserl called this attempt to reduce knowledge to a factual, descriptive, neutral science: *objectivism*. In a decisive reversal, science's human subject became its object, as if the subject and their world were two separate units; as if the social world were independent of the subjects it consisted of: the words used to construct the social world. The consequences of this objectivism are summed up in Husserl's masterly phrase: "Merely fact-minded sciences make merely fact-minded men" (Husserl 1970, 6). The consistency of Kant's distinction between people and things—between dignity and price—vanish: silence in the face of irrationality, injustice and barbarism is the first outcome of this reduction of people to things. The harmful consequences of this apparently neutral attitude—to use Husserl's words, the *mathematization of nature*—are just as visible today as they were a hundred years ago.

The relationship between knowledge and interest, and thereby the denial of epistemological neutrality, is also clearly set out in Habermas's works *Science and Technology as Ideology* and *Knowledge and Human Interests* (Habermas 1971). The task of technical-scientific rationalism is a quest for utility or profitability and the optimization of performance and production. Technical interest seeks mastery: the overcoming of the limits imposed by nature. However, the success of its results swiftly came to be applied to the social world: society and its institutions. As a result, primacy gave way to a form of thinking in which everything was subjected to calculation and planning; where even life, nature and human communication came to be interpreted in terms of the calculation of results. It is the critique of this confusion between technique and practice that has guided critical theory from the outset: "A science that sees the formation of its praxis, which it both serves and is inherent to it, as something that lies beyond it and that is satisfied by the distinction between thinking and acting, is a science that has already renounced humanity" (Horkheimer 1972, 270). Any belief in the emancipatory nature of the technical interest in domination is naïve, and it has disastrous consequences for any technology that claims to be "human-centred".

Objectivism, and subsequently positivism, are the necessary preconditions for understanding the culture underpinning Artificial intelligence and the expectations generated by algorithms. They concur by presenting their results as

neutral and value-free, objective and universal; simply descriptions of reality, akin to a lens capable of seeing and highlighting what is essential—what really exists. Their results are considered to be undoubtedly beneficial for all in the long term, which is an unfounded but effective expectation. In this way, it is quite clear that practical reflection—the question, for example, of an authentic or quality democracy—is meaningless outside the reality that "really exists", i.e., political parties and the aggregation of interests. Democracy now finds itself a prisoner of the methodologies and perspectives of reality that are today integrated into data science, within which the mathematization of reality and the principle of causality yield the central space to the principle of correlation, which is much easier to calculate, measure, predict and therefore control.

As a result, any idea of justice becomes blurred, as the description, recording and datafication of the real turns the generalization of what is found into what should be found. This large-scale naturalistic fallacy has disastrous consequences. All knowledge comes from data science or else it is mere subjectivism. Therefore, out of respect for the equal dignity of all people, the question of justice has been excluded from any possible deliberation and agreement. Given that algorithms define what is desirable by recording and integrating data, the final question to be addressed concerns the location of the autonomy that underlies all forms of democracy, however minimal that may be (Rehg 2015).

Developing a critical perspective provides the means to unmask this algorithmic ideology. Firstly, it reveals the contradiction that scientists commit when they affirm that rationality must be limited to the methodology of mathematics, logic or the datafication of reality, as this affirmation is neither analytical, nor empirically verifiable, nor based on any data whatsoever (Apel 2000). It is simply a valuation; an attitude; an abstraction of the complexity of reality based on the evaluation of science and technology as an intrinsic good. However much we may endeavour to give technical values a preferential place over other values, they are nonetheless values, a decision in favour of a type of action or approach. We are not concerned with a neutral form of assessment, but instead with the imposition of a certain type of value—the value of technology—whose seeming neutrality conceals the power relationships that make it possible and give it power of its own (Fuchs 2017).

There is no neutrality, only a technical concern for dominance, which in this case is articulated through Artificial intelligence and its algorithms. We therefore speak of ideology and the concealment of reality beneath the construction of a new virtual reality, where there is supposed to be no narrative, only the scientific accuracy of data and algorithms. The metaverse and its avatars are the leading

figures of this new reality. From this perspective, when faced by a decision or a clear attitude in favour of a reality, it is easy to perceive how the *neutral attitude of the scientist becomes a specific moral attitude*. The principal contribution made by this critical perspective is to demonstrate that neutrality does not exist. Instead there is a methodological fundamentalism: essentially no methodology is accepted other than that derived via datafication and correlation; in other words, through the application of algorithmic reason. The consequences of this neutrality are clear: without a practical, intersubjectively valid reason, there is no criterion for justice capable of guiding political, economic and social action. The algorithmic ideology can be considered the latest development, so far, of positivism as it considers the knowledge provided by *Big Data*, algorithms and the Internet of Things (IoT) as a faithful description of reality, that is complete and self-contained (Fuchs 2017; Jones 2018). Let us look at a contemporary example of this datapositivism being applied precisely to the language used to present new public policies.

9.2 The Ideology of Algorithmic Neutrality

The following text is taken from the "Artificial Intelligence Strategy of the Valencian Community", an official policy document intended to ensure that the impact of technological development on society should be positive, while also facilitating the coordination of digital innovation policies. It reads:

> The benefits of applying AI will be reflected in better decision-making, as all possible variables will be taken into account and more data cross-checked. Due to the decisions made being based on algorithms they are unaffected by personal subjectivity, which means they are more objective. They are also made faster because computational capacity exceeds the capacity of human intelligence. (Generalitat Valenciana 2018, 5)

This is the Valencian regional government's principal justification for the digitization of its public policies, and it frames all the foreseeable initiatives and innovations to be undertaken by Valencia's administration, companies and universities. It attempts to encompass all areas of reality, providing a clear example of the consequences of algorithmic ideology. Any applications for grants, scholarships, increased budgets, more staff, and so on, must be linked to this worldview; to this way of understanding and explaining the new reality. In other words, it is and will be the main channel for public funding and an

obligatory goal for all financial planning.

This text contains the basic elements of this ideology, and it describes a false reality; a mere abstraction; one which, at the same time, conceals everything that does not concur with it. Let us never forget that the social world, by definition, would not exist without human intervention: it is not independent of the subjects who name and construct it. Democracy depends upon what we understand it to be. Essentially, it is the power of words and of culture: they do not interpret an already existing reality, but construct it.

We will now submit this new version of positivism, objectivism, scientism, or whatever we want to call it to closer scrutiny. Five core concerns underlie algorithmic democracy, which can be summarized as the ideology of neutrality.

The first of these concerns the *benefits of Artificial Intelligence*. Every new technology seeks its ultimate justification in the advantages and benefits it contributes to the improvement of our living conditions. There can be no doubt that technological development has contributed multiple benefits to society. The advances of the internet and digitization in particular have enhanced communication, administration and medical practice, and so on. However, little or nothing is said about the beneficiaries of this innovation are or the negative consequences that have resulted from it. Consider for a moment the initial utopian view of the internet and how it has given way to the current tyranny of the Big Tech companies, as well as the instrumentalization of public opinion, the wide-scale manipulation of elections, and disinformation campaigns, and so on. This technological optimism may be referred to with various names: technical ideology (Wolton 2018), technological solutionism or cyberutopia (Morozov 2013), digital utopia (Pariser 2011), and so on. Regardless of what name we give it, these discourses all contribute to the exaltation of technology as a means of improving and resolving any aspect of social life, but without generating any critical distance towards its effects and consequences. In the absence of any critical perspective, and without any form of participation and deliberation on the consequences of such developments being engaged in by all those potentially involved and affected, autonomy becomes heteronomy or domination. As has been shown in earlier chapters, some governments and platform owners have employed information supplied by internet users for surveillance, manipulating public opinion and as a means to exploit of data for commercial and political purposes.

Needless to say, most of the promised outcomes are long-term ones—curiously always some 20 years away. These deadlines are being extended at the will of those who, whether or not they call themselves scientists, monopolize online media and networks. The promise of greater human and social

understanding prompts us to hope that greater technological development will lead to better human relationships. It is taken for granted that technological progress invariably produces improvements in human communication and that the progress of society depends on technological progress, whether that be in education, politics and culture, or the improvement of the human being in all its dimensions (Larson 2021).

Secondly, the Valencian government's text claims, quite gratuitously, that thanks to AI and algorithms, progress will be made in *improving all decision-making*. We have already discussed in the first part of this book how democracy should not be identified as a representative system, which is the view set out in most new democratic revisionism. For a government to be democratic, elections do not suffice. They are a necessary condition but not a sufficient one. The same applies to decision-making, which is undoubtedly a key moment in the democratic process. Technocrats can learn to use elections to their advantage, as we shall see when discussing post-truth. They can also present themselves as the most capable decision-makers, but they cannot drain democracy of all meaning, which, let us not forget, has to do with the *inclusion of* all those affected by decision-making and, likewise, consideration of them when calculating any consequences.

Democratic legitimacy, however much it may be reduced, cannot eliminate the participation of those who will suffer the consequences of the legally binding decisions taken, or the creation of public policies. Reducing democracy to politics, and in turn to elections and expert decision-making, is mere electoral fundamentalism, and it blurs the responsibility that should be taken by civil society through its confusion of the political and the public (Mair 2013). To reduce the people's participation to the election of representatives by means of voting is to remove citizens' autonomy (Cortina 1993).

However, the age of innocence is over for algorithmic colonization, as its effects have become all too visible. European guidelines, along with UNESCO, warn us of the dangers of using algorithms in decision-making. For example, due to inherent biases, algorithmic decision-making can implement or exacerbate: discrimination, inequality, digital segregation and exclusion, a threat to cultural, social and biological diversity, social and economic division. Transparency and an understanding of algorithms and the data they are trained with is therefore necessary. There is also a need for greater awareness about their potential impact on *inter alia*, human dignity, human rights and fundamental forms of freedom, gender equality, democracy, social, economic, political and cultural processes, scientific and engineering practices, animal welfare and the environment and ecosystems. (Mittelstadt and Floridi 2016; Mittelstadt et al. 2016). None of this

appears in the documents that provide support for the digitization of our society, nor in the criteria underpinning the distribution of digital resources.

Decision-making is, of course, important in public policy, but, as with the tools of Artificial intelligence, we must establish who participates in it, what criteria are used to assess and evaluate it, and who defines these criteria? Who forecasts the consequences, and from where and how do they collect the data? Moreover, what is all this information checked with? We are not just agreeing that radar is the best way to steer the ship, we are raising the issue of whether it has occurred to anyone to ask what port we want to get to, which way we want to go, and how we want to live together on the way. It seems that we must entrust the answer to algorithms.

Thirdly, there is *blind trust in algorithms*. Data is the outcome of algorithms and, in turn, their staple source of nourishment. Data describes empirical facts that are organized or tabulated by algorithms. When we talk about data, we are talking about specific information, collected as numbers, letters, symbols, and so on that are themselves meaningless, and have no content or significance. By being processed for data collection and synthesis by algorithms, models or what might be called criteria it becomes valuable information that can be used for a range of specific purposes. However, this process of selecting and collecting fragments of reality and our own identity for uses that are, in principle, very often unknown is neither neutral, nor free of risks and dangers (Tsamados et al. 2022).

Algorithms, in turn, constitute an ordered and finite set of operations for finding the solution to a problem. They provide an abstract and effective control structure that fulfils a specific purpose under certain conditions (Hill 2016). They are a form of software code that processes a limited set of instructions; essentially, an algorithm is any mathematical implementation within a form of technology that is applied to undertake a task and achieve a result. If data is integrated into or created by algorithms and we are firmly confident in their objectivity and neutrality, it is clear that more and more decision-making processes will use them because they grant them greater *legitimacy*. As a result, algorithms are protected from any users' suspecting that their criteria are biased towards the commercial or political benefit sought by the provider, and their use of embedded and unexamined assumptions that operate below the level of awareness of even the designers is similarly concealed (Gillespie 2012). How can we trust what we do not know, whose criteria for quality, or even justice, are unknown? How have we reached the naïve belief that data and algorithms are truly neutral?

This is fertile ground from which we must forge a critical perspective

capable of making visible what would otherwise be ignored due to dogmatism and a technical interest in domination. There is a framework of meaning—a reduction of reality to the measurable and quantifiable—that assumes a prior interest in objectification and mathematization. This assumption is guided by specific values that are far from neutral.

The fourth concern raised by the Valencian government's statement is that this power of neutrality is exposed by its claim that algorithms are not affected by *personal subjectivity*. The neutrality of analysis is the naïve view of the intentionality concealed by the abstraction of the classification and learning criteria for the data and algorithms' classification. This is the linchpin for the construction of this algorithmic ideology, according to which the incorporation of technological improvements, backed up by calculations, will improve any aspect or area of social reality. If a human activity can be measured, then it can be calculated and predicted; if it can be calculated and predicted, then it can be optimized and improved. If the world is imperfect, it is because humans are imperfect. In contrast, algorithms do not make mistakes; they undertake precise calculations, and consequently their results and predictions can only yield the best results (O'Neil 2016).

The question is how this belief, which makes us expect more from machines and trust them more than people, has been developed. How has this favourable and unconditional attitude; this optimism and blind trust in digital technology as a driving force for development and, of course, justice, come to be established? Our refusal to recognize any human intervention, as if the data were "out there" waiting to be discovered; as if the algorithms were not the *de facto* outcome of our interpretation of reality—our interested pursuit of knowledge—not only demonstrates the clear short-sightedness of this attempt to understand the situation, also it also shows a clear intention to *reify,* and thus *obstruct* any possibility for ethical design and governance (García-Marzá 2005). If algorithms merely describe and collect what "exists", they tell us nothing about what "should exist". Making this case is the first task of a dialogic digital ethic for conceiving of alternatives to algorithmic democracy and its derivatives. However, as will be seen below, it is not the only one.

The fifth and final concern raised concerns the conclusion drawn on the basis of this ideology of neutrality, which leads us to the unquestioned need to *replace our ability to decide and act*, in whole or in part, with algorithms. The Valencian government's text makes it clear that it views Artificial intelligence as superior to human nature (a concept that is neither defined nor justified anywhere). By human nature, we are referring to the limits of human capacities and competences to calculate, measure and predict, making available everything that

exists. We seemingly have to resign ourselves to the fact that our capacity to undertake calculations has long since been overtaken by algorithms. Incidentally, this is a claim that no one denies. However, a very different question is posed by the fact that the belief that both reasoning and arguing in order to convince are the same as calculating (incidentally this is wholly unscientific positivism) has now become hegemonic. A simple critical overview enables us to explain the main characteristics of the algorithmization of the real, and to show how these characteristics are neither independent nor neutral in terms of their design, development and prediction of results. Instead, they depend on the interest invested by programmers or developers, and the technological companies that use them, as well as governments, in addition to the possible commercialization of their results by markets and states.

Firstly, algorithms are fed with data (input and output). But untreated raw data does not exist. It is not “out there” waiting for us. Data is an abstraction that assumes an intention; an interest. It is constructed. The “original” data is already an interpretation of reality, responding to the concerns that underpin our knowledge, and providing the framework with which we simplify reality by eliminating from it whatever does not interest us. Data always consists of interpreted, evaluated facts. Secondly, decisions made by algorithms may be incomplete and therefore produce unreliable data, and they may be biased or simply wrong. Biases can arise from the pre-existing social values encountered in the social institutions, practices and attitudes that produce technology; from the people who handle the data at source; and, of course, from the interests that are expected to be met, as well as the kinds of preference integrated into the algorithms. Consideration must also be given to their complexity and capacity for self-learning, which converts them into real “black boxes“.it is hard to discern who has intervened and how in the creation of algorithms, and likewise it is difficult to detect how great an influence the inevitable human subjectivity has had on the design and configuration of algorithms. The results obtained using an algorithm require interpretation: their meaning is not self-evident from their statistical models; they must be interpreted and someone must define the criteria for any evaluation of their results. The answer to the question of whether an algorithm is good or bad is not merely a practical one, as it must also include the position of those who will suffer its consequences.

9.3 The Construction of Algorithmic Culture

Ideology must not be confused with culture, even though they share the same semantic field. The former refers, more negatively, to the *instrumentalization* of

culture for attaining a specific reality and interest. In this case, we are talking about the blind and unfounded belief that the results of datafication should be broadly applied, not only with regard to overcoming the limitations nature imposes upon us, but also with regard to domination by one social group throughout society and its exploitation of another group. Such a scenario is the outcome of technical rationality focused on effectiveness and efficiency becoming confused with practical rationality, which is centred on justice.

In a broad sense, ideology is directly related to a general idea of culture as a network of ideas, beliefs, senses and meanings that assign values to life and to social practices. As a means of instrumentation, ideology gives rise to a reorientation of itself: it uses power to deceive and manipulate. Algorithmic ideology requires a broader set of ideas, values, beliefs, and so on, as well as signifying practices and symbolic processes to imbue it with meaning and credibility. Within this framework of the construction of meaning, it becomes clear how algorithms are, and can be nothing other than, generators of culture: a cosmovision according to which they are the centre of social reality and of any possible means of improving it (Finn 2018).

An example of this algorithmic culture is the value placed on neutrality, which is by no means neutral. The reason for this insistence on algorithmic neutrality is plain to see; we intuitively associate neutrality with concepts such as objectivity, asepsis, equanimity, impartiality, independence, rectitude, fairness, integrity and, of course, justice. It is this semantic shift that enables us to speak of consistency and legitimacy, that is, of a belief in justice accompanying algorithms. Algorithmic culture is responsible for justifying this shift from objectivity to justice. It is algorithmic culture that struggles for hegemony and today underlies and supports algorithmic democracy.

Defining, maintaining and developing this way of seeing reality and of making sense of what we do or do not do with AI is the basic task of a culture that sees technological advances as the solution to all of humanity's problems. It explains how this belief has permitted computer algorithms to shape and condition users' activities and decisions, both in the online and offline worlds, thereby shaping what has come to be understood as algorithmic culture (Gillespie 2016).

When we talk about algorithmic culture, we are not referring to a new culture emerging in the digital world or in cyberspace. Instead, we are talking about the authority granted to the algorithms that conditions or even determines the senses and meanings we construct and that "profoundly shape the cultural experiences of people's activity on social media platforms" (Van Dijck 2016, 32). By using this concept, our concern is to highlight algorithms' capacity to "create" culture,

incorporating the forms of behaviour, lifestyle, sense and meaning that all form part of the vision of the world provided by Artificial intelligence and its technologies into our way of feeling and acting. What we term algorithmic culture refers to the ways in which our cultural experiences are increasingly determined by algorithms. It is therefore imperative to stop viewing the latter as neutral, objective or aproblematic; the decisions they make come to shape and predetermine our online and offline behaviour, lifestyles, customs, worldviews, relationships with our peers and others. For this reason, Finn argues that it is possible that computational algorithms are presented as being simply mathematical, but what is clear is that they behave like cultural machines that drastically revise the geography of human reflexivity (Finn 2018, 96).

This culture and this way of thinking, talking and doing things, has led to a kind of algorithmization of social life, which some authors call *dataism* in order to identify the belief that only the measurability and quantifiability of reality matters, as well as the high expectations of improvement of this reductionism (2016). As was discussed above, this simplification of reality gives rise to the reduction of people, society and the world to quantifiable, verifiable and calculable data that can be manipulated by programmers seeking certain forms of behaviour. These indirectly shape our manners of thinking and acting. Both authors believe this conception of the digital age becomes deeply contradictory, as it apparently claims to overcome all kinds of ideologies, but it is itself an ideology, due to the series of beliefs and dogmas it generates about our hopes and desires, which are impossible to ascertain. In short, the characteristics of this ideology of dataism is a widespread belief in the virtues of objective quantification and the potential tracking of all kinds of human behaviour through digital technologies. This belief has now been incorporated into our way of seeing and conceiving of democracy and its possibilities. It has become a machine for creating meaning and, therefore, reality.

Culture, with its value systems and rules, is not the paint we apply to complete the construction of a house called democracy, but rather the foundation upon which it is built. Records and data storage are not used for storing meaning, nor the reality we feel comfortable and safe within, and which we refer to as normality. Instead, meaning is contained in everyday language—in the narratives used to interpret symbols and data. A good example of this construction of what is deemed normal—what is expected and desirable—is the alienation of those who, because of their age or knowledge, are considered to be *digital emigrants*. Today this isolation is akin to a new form of disability and a feeling of guilt for not being able to keep up with digital technology. From this perspective we can now understand statements such as: that there are no good and bad algorithms,

solely the good use of them. Once again, neutrality has been applied across the board (Striphas 2015). Blind trust in Artificial intelligence technologies, their contribution to the future of humanity, and the consequent neglect and abandonment of any critical perspective, is the visible face of this culture. In short, the technological revolution can only appear when it has already taken possession of the lifeworld (Kitchin 2017).

The strength of this algorithmic culture is based on the expectations created by the promised improvement expected to be derived from its application. Algorithms have become meaningful in the cultural imagination; they are "accepted" not because they refer to something transcendent in the classical sense, but because they have attained a form of acceptability in a newer and more immanent way. Scott Lash's idea of the principle of the "legitimization through actuation" of algorithms is fundamental in this respect (2007, 67). Through their real-time deployment, algorithms implicitly or explicitly claim to be cost-effective and objective, both in an epistemological and moral sense. Algorithms work because they offer solutions—they allow us to satisfy desires and interests—which are usually constructed by other algorithms. In this sense they are generators of trust: practical and symbolic guarantees that their assessments are fair and accurate, free from subjectivity, error or attempts at influence (Gillespie 2016, 179; Mager 2012).

It is indisputable that today our behaviour and use of social media along with the internet determine cultural practices. As part of culture, algorithmic culture is inscribed within power relations and interests that generate new inequalities and forms of discrimination: it is not an idyllic and aproblematic culture as those who own it claim. In the case of algorithmic culture, the interests of the platforms and advertisers—the intermediaries who own the algorithms that collect and process our data—that come into play. The recommendations made by the algorithms are based on the information they collect and correlate, so their recommendations and predictions are not the same for all users, as they depend on many factors and variables, some of which can be discriminatory, such as location, income, age, ideology or sexual orientation.

Virginia Eubanks's *The Automation of Inequality* presents a range of examples of discrimination concealed by the supposed asepsis of algorithms when used for decision-making as part of the establishment and application of public policies. Given, as we have shown, that we are dealing with truly cultural machines, it is easy to deduce that, even if we are talking about causalities, on the basis of correlations of a definitive description of an existing reality, the justice of this society cannot be extracted mathematically. However, Eubanks shows that the use of algorithms perpetuates inequalities and always punishes

the same people: austerity and marginalization for those who have the least. We are not dealing with neutral, apolitical tools, but with the application of certain value-driven models that select what they expect to find. Yet technology, as we discuss in more depth below, allows us to dilute our responsibility. It offers us an excuse to do nothing to solve critical problems that require a sense of justice. The author tells us that the algorithm is, in the end, a way of exempting us from the human consequences of political decisions. (Eubanks 2018; O'Neil 2016). Essentially, it allows us to shirk all responsibility (Habermas 2021).

According to Antonio Gramsci, hegemony is the ability to bestow meaning on what we do, say and expect from others. Despite this situation, human beings, consciously or not, have come to hold an unequivocal belief in algorithms, without pausing to reflect on the implications they have for our lives; our ways of seeing and understanding the world around us; and of relating to one another. Thanks to their efficiency and effectiveness, we have come to rely on algorithms without questioning their results. However, a range of authors have begun to draw attention to their impact (Finn 2018; O'Neil 2016; Pariser 2011). We may think that they make life easier for us to make decisions, but they are giving rise to new processes of discrimination and inequality and these contradict the democratizing and equalizing aspiration promised by the evangelists of the digital world. In this sense, it is worth asking ourselves whether we should continue to blindly trust the decisions algorithms make on our behalf, while ignoring that we are dealing with a form of power that manipulates public opinion and the formation of desires and preferences. This is the core foundation of algorithmic democracy, which enables it to construct an invented reality and give it meaning. It is the basis of its power. How can this power be concealed beneath a cloak of unbridled enthusiasm and naïve optimism? One possible answer can be found in a natural ally of algorithmic culture, not to say one of its basic assumptions: the emotivist concept of politics derived from neuropolitics and neuropower, which have today been integrated into the pretensions of algorithmic democracy. The combination of these two elements explains the strength of this new ideology: its hegemony.

9.4 Neuropower: From Neuropolitics to Neurodata

The time has come to develop an understanding of the kind of power we are dealing with when talking about algorithmic democracy. The strength with which this ideology has taken root in political thought and the future prospects for democracy cannot be explained without reference to its deployment of two key tools, neuropolitics and post-truth, whose interests converge. Populism has

been able to exploit both the increased role given to the emotions in politics, and the disregard for any shared reference to public reason, truth and justice. The result has not been an expertocracy, but rather autocracy, as the Russian invasion of Ukraine has sadly made apparent.

Democratic theory has always sought reasons to explain these forms of disaffection with regard to the social and political situation that leads to it. We reviewed these forms of disaffection in the opening chapter: a failure of the so-called welfare states to meet expectations; the lack of political representation and political parties' impotence when confronted by a globalized economy; the conversion of politicians into a class or caste; the lack of political programmes or any intention to fulfil them, and so on. If we attend to the intuitive knowledge that we possess as citizens, as members of a democratic state under the rule of law, we can discern how the meaning of democracy is currently being undermined: any expectation that a form of reason can be attained through political autonomy has been lost. If the democratic principle refers to the fact that all those affected by laws and public policies must have given their consent to them—that any citizen affected by a decision must have the same possibility of influencing it—this horizon of meaning no longer exists for what citizens understand, and therefore expect, of their democratic system.

As we discuss in these final chapters, with regard to what are known as deliberative democracies the solution would involve, firstly, the necessary reform of democratic institutions, political parties and systems of representation. Then would come the empowerment of citizens, the strengthening of mechanisms for participation and greater control exerted by civil society. This two-way democracy therefore requires a return to a radical vision of the democratic principle because it continues to insist on free and equal participation. This is the only way to respond to a form of power, which is defined by both the possibility of satisfying interests and the ability to define them. The political process is thus understood in terms of deliberation, the exchange of arguments and the search for rational agreements or, failing that, compromise and fair negotiations (Habermas 2018). Public debate is necessary so that every citizen can provide the reasons needed to legitimize actions, rules and institutions. Whether for strategic reasons or moral reasons, there is no other way to talk about legitimacy unless we "disconnect" decision-making and power from the rational discussion that justifies them. This is what neuropolitics contributes to algorithmic democracy.

When we speak of power, we generally speak of the power to shape and change the reality in which we move: the power to make decisions and to do what we want or should do (Wrong 1997). Since the emergence of classical

sociology, this concept has always been related to the existence of conflicts; to the ability to pursue one's own interests in the face of or against others, i.e., in spite of resistance from others. In other words, it concerns the ability to influence others and make them act differently from the way they would like to act. In short, Weber refers to power as the probability of imposing their own will within a social relationship, even against all forms of resistance and irrespective of the foundation for this probability (Weber 2019).

However, this concept of power is insufficient when it comes to defining the values that make up algorithmic culture for one simple reason: the existence of conflict is not the only arena in which power is developed. It overlooks the fact that power exists not only when we are able to resolve a conflict of action, but also—and above all—when we manage to prevent such conflict from arising. In complex scenarios such as today's, power has become a social relationship that is barely observable or invisible: it functions by articulating and selecting social conflicts, avoiding decision-making through the creation of values and institutional practices. Whoever has the power to construct "social meaning" can prevent the interests of other groups being organized or articulated. They thereby remain latent or as sidelined interests. It is even possible for certain groups to persuade other groups to take on the interests and expectations they consider to be in their own best interest, expressing the needs and desires that suit those with power (Lukes 2021).

The basic characteristic of this type of power is its invisibility, as individual preferences, desires and thoughts become open to external managers. Control over the education system is now being incorporated into datafication and, with it, the commercialization and control of our own data, with which personalization and microtargeting will become possible. Power is also expressed when we *decide* what data is, when we *construct* the algorithms that make sense of Big Data, when we *assess* what is a good result, and so on. In addition to the fact that we are leaving traces all over cyberspace, neuroscience —in this case neuropolitics—is currently informing us how we can create new stimuli and new forms of intervention in our consciousness and will that depend on whoever is in control, whether they be commercial or political interests. The relationship between neuropower and algorithms demands closer scrutiny.

Neuropolitics has become a pioneering neurodisciplines. It is defined as the study of the neural correlates underpinning political decisions: citizens' decision-making and behaviour, the election of representatives, political participation, electoral behaviour, the affection or disaffection provoked by political leaders, and so on. Essentially, it is concerned with understanding people's brains with regard to their being citizens. To achieve this aim, it is merely necessary to

switch on the scanner and observe what happens. Two surprises then appear.

The first concerns what we could call the myth of *homo politicus*; the idea that citizens' ability to choose rationally between competing political candidates, as well as monitor their actions and assess whether they deserve the renewal of our trust is wholly unrealistic. It does not exist. The use of reason in politics, whether for deliberation or for calculating how we will fare under each candidate, is not the fundamental building block for making choices. On the contrary, when we trace the movement of our neurons, we realize that many of our beliefs are intuitive, many of our responses are automatic and most of our decisions are unconscious, to the extent that on many occasions we do not even know why we have voted or chosen this or that option. The citizen whose actions in the sphere of politics, as in that of economics, are based on their calculations of the costs and benefits of their choice, does not exist. However, it would be another matter if we could access this inner world and datify it (Conill 2019).

The second surprise concerns the emotions, a well-known theme in democratic theory. The workings of the brain reveal to us that emotions do not play a secondary role in decision-making: they are not a mere accompaniment. Instead, they are arguments that must be explained on the basis of our emotional capacity. Emotion should not be seen as a failure of reason. On the contrary, considering it enables us to open up potential for social change and transformation that can now be explored and measured. Our conception of politics and, in short, of what we expect from democracy, changes radically if we start from the assumption that our political decisions are the product of emotions and not reason; that our political brain is, by definition, an *emotional brain*" (Westen 2007; García-Marzá 2013).

Lakoff was among the first to assert, as the subtitle of his book *the political mind* states that "… you cannot understand 21st-Century politics with an 18th-Century brain". We are facing what he calls "the trap of rationalism", a form of understanding political life inherited from the Enlightenment and its principal motto: "the truth will set us free". According to this belief, all human beings are rational and reason is something that is objective, universal and free of emotion. Therefore, if we reason together we will achieve political truths and thus solve problems. Hence, we can govern ourselves, without the need for kings or warlords. Given that we are equal in reason, the best government is democracy. While this continues to be the textbook understanding of democracy, today we know from cognitive science that it is divorced from reality, Lakoff informs us, that people think in frames, and each frame—for example, strict vs. protective parenting—imposes a certain logic. In order to be accepted, truth has to fit into

the mental frames people use. If the facts do not fit into certain frames, the frame remains unchanged and the facts bounce back (Lakoff 2008).

A detailed discussion of Lakoff's theory cannot be pursued here, but it should be mentioned that for this author, frames are mental structures that shape the way we see the world: our identity, our goals and objectives, our way of thinking and acting, our assessment of what we believe to be good or bad, fair or unfair, and so on. However—and here we begin to enter the realm of neuropolitics—these mental structures are neural networks of association that can be accessed by language through metaphors, under which the use of frames becomes unconscious and automatic. We use them without realizing it, because they form part of the collective unconscious and are physically embedded in the brain's synapses. When we vote, we do not do so by checking whether the candidate's programme matches our interests, but instead because our decision depends on whether the candidate has been able to activate the frames that are already physically present in our brain in the form of neural networks. This is usually achieved through repetition, metaphors, narratives and symbols (Lakoff 2014).

However, neuropolitics can no longer give a reason why some values carry greater weight than others. Values that are presented as progressive—e.g., trust, sincerity, a strong nation, prosperity, a better future, effective government, shared responsibility—have no specific content outside the context in which they are used. As José Ortega y Gasset said, it is one thing to make continuous repetitions until values are framed, and quite another to know whether or not these values are meeting the moral standards of our times.

These influences have begun to have a strong impact on democratic theories. The study of emotions can offer an alternative framework to explain and predict political choice and action. Once more, the starting point is a familiar question: what if we were wrong about the overall impact of the emotions on decision-making? What if emotions in most instances performed a productive function, providing the basis for quick and accurate decisions? In short, what if emotions assumed an equal or even primary status in decision-making?

It is precisely the vindication of emotions that is the basic pillar of populism, as majorities are formed by manipulating citizens' feelings. Political elites manipulate the voters' feelings and emotions to win their votes or approval for their policies. Fear and indignation, for example, instead of being understood as reactions of powerlessness in the face of specific clearly unjust material conditions of life, are converted into misalignments of the brain's electrochemical mechanics. Essentially, what has fallen by the wayside is our ability to guide our own life as a community, as well as our accountability and

our political autonomy, which leaves the door wide open to populism. The interrelationship between neuropolitics and neurodata becomes wholly apparent through a consideration of how neuropower makes it possible for this manipulation not only to be permitted but to be welcomed with open arms (O'Neil 2016).

When we speak of neuropower, we speak of a form of power exercised through the mental modulation of individuals and aggregates' processes, functions and expressions. Through brain imaging, medical technologies and treatments, and by offering conceptual paradigms, neuropower acts on mental functions that were previously invisible, ignored and unmanageable. It forms part of the regulation of cognition, sensation, attention, mood and, for the most part, emotions. In short, neuropower is a form of power that operates through the mental modulation of individuals and aggregates' processes, functions and expressions (Dunagan 2004).

The determination and speed with which neuropolitics is developing cannot be fully gauged if we do not consider the most important issue: the technical and practical application of the results achieved, i.e., the political and economic use of the advances made by neuroimaging techniques. The brain-society interrelationship has a direct technological application and this seems to be the reason for so much rhetoric about objectivity and science: measuring, quantifying and managing—in other words, controlling and predicting. As has become apparent, the time periods referred to as "eras" have shortened. We have moved very quickly from the era of neuroscience to the era of data science.

Dunagan's work provides a valuable description of the forms of power emerging from neurotechnologies and their integration with the current Artificial intelligence technology. We will now turn to examine two examples of the interrelationship between neuropower and what has been discussed above as algorithmic democracy.

The first example is the new neuroimaging technology that produces mountains of new data and can be used for surveillance, detection and control of undesirable behaviour, regardless of how this concept may be understood. This form of measurement can be used for the construction of sophisticated lie detectors, when applying for a driver's licence to determine any predisposition to suicidal driving, as well as for detective work. Governments have become obliged to develop these techniques to ensure the security of their citizens, and because they can prove effective in implementing public policies.

Another example of the integration of neuropower and algorithmic democracy are techniques of standardization and improvement. Thanks to computer processing—the true neural grammar—neural data is being

increasingly quantified and is thus generating possible standards for memory capacity and a maximum concentration spectrum, as well as for what have until now been subjective aspects, such as happiness or well-being. Once we have accepted the criteria of normality, the next step is to try to improve our mental effectiveness. Why should we not try to improve our mental performance ratios, in terms of emotions and thoughts, with pharmacological means such as oxytocin—the hormone responsible for empathy and trust? Or with magnetic devices that stimulate or inhibit certain areas of the brain? Or even with neural prostheses, which are made possible thanks to advances in nanotechnology? Or by fitting a chip capable of processing more information faster?

However, perhaps the most recent, and most dangerous, example of this integration is *neurodata*: the result of applying so-called data science to the images offered by neuroscience. *Brain-computer interfaces* are the real possibility of connecting our brains directly to the network, needless to say, through algorithms. They have already been developed, which has caused alarm among scientists (Yuste et al. 2017; Wolkeinstein et al. 2018). Of course, public opinion is only presented with the good results that are foreseen. By being able to "read" brain activity and even engage with it, we will be able to better understand how our minds work, help the sick, and create economic opportunities. Yet, what if someone could access our subconscious without authorization? How could we control the information in our brains if connected to networks belonging to large technological platforms or states?

In short, neuropolitics means politics seems to have become detached from the idea that defined it; as Habermas put it: the formation of opinion and common will. Is this a cause or an effect? It is better to speak of complementarity and feedback between a political situation that lost control in terms of democratic management and, so to speak, the neuroculture, which conjures up the fascination and magic surrounding neurotechnological advances that promise to solve these imbalances by improving—or rather manipulating—citizens' brains. With this, what are clearly situations of injustice become misalignments of the brain's electrochemical mechanics. Instead of improving the social and political world, neuropolitics invites us to improve our neural systems. Neural correlates are converted into authentic narratives of legitimation of clearly unacceptable and unjust situations. Once datified, the metaverse will provide us with the alternative reality we seek, and can afford.

What lies behind these expectations is the same scientistic reductionism of confusing cause with motive, as if neuronal movement were the cause behind decision-making. However, this determinism, which attributes mental events to actions caused in the neuronal synapse, is out of touch with reality. To be more

precise, it lies beyond social, ethical, political or economic reality, as this reality is not independent of the subjects who construct it, however institutionalized it may be. It is demonstrated by a persistently recurring fact that is impossible to overturn no matter how much neuropolitics is introduced: this social order depends on the validity granted to it by the subjects, i.e., it requires legitimacy. Democratic practice not only presupposes emotions, it also requires knowledge that permits us to distinguish unjust situations and to adduce the reasons we use for our assessment. But to speak of justice requires a shared reference to a reality that we believe "ought to be". Is it possible for democracy to have an epistemological value—a reference to practical truth—that enables us to speak of justice beyond, and independently of, datatification and algorithmic colonization?

To answer this question in an affirmative way necessarily implies the that the concept of reason is intersubjective and can be made universal. What characterizes practical reason is precisely the ability to justify our actions and institutions with reasons that must be valid for every rational subject—every subject thought of as a valid interlocutor. In the *Critique of Pure Reason*, Kant stated that this public character of reason is a dimension that is not given or present "out there", but rather must be constructed among everyone. This tends towards a possible agreement among everyone:

> Reason must be subject, in all its operations, to criticism, which must always be permitted to exercise its functions without restraint; otherwise its interests are imperilled and its influence obnoxious to suspicion (…) The very existence of reason depends upon this freedom; for the voice of reason is not that of a dictatorial and despotic power, it is rather like the vote of the citizens of a free state, every member of which must have the privilege of giving free expression to his doubts, and possess even the right of veto. (A 738/766)

Kant went on to insist: "We have then the means, though only subjective means, not, indeed, of producing conviction (Überzeugung), but of detecting the merely private validity of the judgement; in other words, of discovering that there is in it the element of mere persuasión (Überredung)." (A 820/B 848). As we shall see, these ideas will give way to the principle of explicability, which is a key principle for a digital ethic and, therefore, for a democracy capable of curbing the current algorithmic colonization. But there is also a negative response, which complements algorithmic culture and neuropower: the affirmation that truth does not exist and that only my truth exists, thereby

diluting any reference to a just society, the ultimate goal of democracy.

9.5 Post-truth: The Power of Technopopulism

Disinformation is one of the most significant social pathologies our societies suffer from, affecting the formation of our identity, beliefs and interests, as well as our democracy as a system of institutions depending on the possibility of public conversation and debate. At stake is not only our freedom, but also our sense of justice, for example, the meaning of what we have hitherto understood as a decent life. To deny that there is truth is to deny that there is any possibility of criticizing the injustice committed by our institutions, including their democratic decline. Indifference to reality is also indifference to injustice; it is a refusal to see the inequality corrupting our democracies. On the one hand, the increasing inequality has given way to populism and, with it, the replacement of reason with emotion. Meanwhile, the potential of digital technology has been put at the service of mercantilism and partisanship. Both developments underpin the current post-truth system (García-Marzá 2021).

Disinformation is one of the most important social pathologies we suffer today and a breeding ground for the advance of algorithmic culture, as was mentioned above when we discussed the demise of public opinion. The concept of public reason should not be confused with that of public opinion, which is much broader. It contains the reasons that justify or legitimize all kinds of power, including, of course, the power embodied in democratic systems.

In 2016 the Oxford Dictionary chose the neologism “post-truth” as word of the year. Its definition refers to a context or situation where objective facts are less influential in shaping public opinion than appeals to emotion and personal belief. The concept is intended to describe a new reality that has little to do with the truth and a great deal to do with what we believe and feel. The concept is not an attempt to describe manipulation, lies and deception, which have always been part of political strategy. Instead, we are facing a real social transformation; a paradigm shift where “truth” no longer seeks any correspondence with reality, but describes only personal attitudes, emotions and feelings. *My* truth becomes *the* truth if I have the power to make others believe my message and feel it. While in 2016 it was Brexit and the election of Trump, the renunciation of the quality of public opinion and its value as a basic means of shaping an individual and collective will has now become a normal situation that serves to conceal the injustices undermining our democratic systems. This is the new normal (García-Marzá 2021).

Indeed, the World Health Organization already uses the concept of *infodemic*

to refer to the excess of false information, hoaxes and rumours which hinders people in accessing reliable information for making decisions when faced by a pandemic. Disinformation, they say, has become the coronavirus's best ally. The same is true in the political arena, in elections, and when it comes to gauging the efficiency of public policies. There are two ways to overcome injustice: to try to transform reality, or to construct a parallel reality which, by relying on beliefs and emotions—in other words, by preventing debate—manages to hide injustice, making it invisible. Post-truth works in the latter way and is a deliberate distortion of reality aimed at manipulating beliefs and emotions in order to influence public opinion and social attitudes (Harsin 2015). In line with this definition, post-truth is a way of narrating reality based on our own desires and interests—our own subjectivity—in order to induce behaviour, change attitudes and seek complicity. Any reference to reality is subsumed within a narrative in which emotions or opinions carry more weight than factual data. Its purpose is the same as that of marketing and advertising: to influence. In short, post-truth is a deliberate distortion of reality intended to legitimize a certain position of power, be it political, economic or social. It renounces debate and, with it, any accountability that does not rely on emotional narratives and is clearly removed from either theoretical or practical reasoning. "Truth" is no longer a criterion of validity, but rather an attitude; a feeling.

Post-truth blurs the boundary between truth and lies, creating a third category which is different from the previous two. This is one in which a "fact", which may or may not be fictitious, is accepted in advance simply because it fits our mental schemas. False information becomes *alternative facts*. There is no one truth, only different interpretations of reality that are equally valid. The winner is the one who manages to mould their truth around our feelings, beliefs and desires. Neuropolitics provides us with the ostensibly scientific basis for this embedding of frames of reference. Inequality is therefore no longer an injustice, it is an opportunity to innovate and progress.

The novelty of this new scenario is not only the continued and widespread use of fake news and our rulers' increasing mendacity, but also the *public response* (D'Ancona 2017). Indignation gives way to indifference and, finally, as happens with climate change, alternative truth enters the picture. The world's scientific community has been warning about global warming for decades and has proven it with data and studies. However, alternative truth says that climate change is not so serious and is apparently caused by natural phenomena. It is easy to predict which is more comfortable to believe when faced with the false dilemma between economic development and respect for the environment, and all the more so when a person's standard of living, employment and family

depend on this decision.

The source of the radical change of scenario represented by post-truth is that the institutions that structure any democracy have failed to meet the needs legitimately expected of them. In addition to the corruption of our rulers and the business sector, there is the growing inequality that democracy itself has generated and maintained. Citizens' disaffection is not the result of apathy or disinterest, but rather the systematic failure to deliver on the promises that provide the confidence and lifeblood for our democracies. Democracy still retains its aura of legitimacy, but it is no longer the only system that dares to present itself in public. Yet there is talk of democratic authoritarianism or illiberal democracy—systems that rely directly on populism and its core tool: post-truth.

If the origins of post-truth lie in this concealment of inequality and the populism it gives rise to, the driving force for its development is bound up with the new digital technologies, the internet and social networks. We have already discussed the disappointment resulting from seeing the internet go from being an optimistic tool for free, equal and widespread communication to being a space occupied by a few digital-technological multinationals, which operate with a clear monopoly and have commodified and politicized any sense of expectation. Disinformation is closely linked to the digital revolution that constitutes the ideal infrastructure for fake news, hoaxes and canards for this *post-truth regime*, as Harsin (2015) refers to it. Populism needs this dependence of truth on power—the so-called alternative facts—and, in turn, this dependence is built upon the personalization provided by the new digital technologies, which according to Pariser create what are known as *bubble filters* (2011). If democracy implies dialogue and taking into account the point of view of others, this is prevented by directing biased information towards citizens who are separated and isolated from each other: information that only confirms their own beliefs. Immersed in the solitude and individualization resulting paradoxically from "social" media, we become easy prey for manipulation as we are no longer able to distinguish clearly between truth and falsehood, between justice and oppression, between respect and contempt, or between sincerity and pretence.

It would not be possible to understand this difference without being clear that reality exists and truth is intersubjective, a claim to validity that we must define together, through dialogue and agreement. This agreement implies a moral core, a demand for universality, a reciprocal recognition that allows us to speak of harm, falsehood, intention, damage, and so on. This is a criterion that post-truth tries to bury, leaving us without reasons for understanding reality and jointly seeking possibilities for change and social transformation. Trust is only possible

if we have reasons intended to be applied to everyone in order to expect a certain individual behaviour or response from an institution. Although they have a cognitive element, emotional reactions clearly do not suffice to maintain any kind of interaction or shape joint expectations.

Thus, in summarized form we have set out the three dimensions upon which the false idea of algorithmic reason is based and, therefore, the culture that surrounds and legitimizes the possibility of algorithmic democracy: the ideology of neutrality, the predominance of emotions in politics and, finally, the blurred pretence of justice that accompanies the algorithmic reduction of social, political and economic reality. This time it is not about the risks and challenges to be faced over the next 20 years, but rather a current reality making it possible for the different forms of revisionism discussed in the first part of this book to form an authentic anti-democratic front: *technopopulism.* This describes the application of digital technologies for populist ends creating an integration between technocracy and populism, giving rise to an expertocracy or epistocracy, as it may be termed, based on the knowledge provided by marco data and its use for the formation of power. In short, this provides a new logic for democratic politics. The message is clear, going back to a utilitarian justification: what is important is to find solutions for the things that really matter to people (Mair 2013).

The term technopopulism refers to politicians and their allies who have learned to manipulate social media—Cambridge Analytica, for example—using its algorithms to incentivize their followers and influence those who are dubious and vulnerable with the aim of controlling narratives, attacking the press, increasing polarization and destroying reputations. The disinformation strategies used by these digital populists are necessary to discredit and delegitimize both the incompetence and corruption of the political elites and the servitude of the professional press. Today it is easier to fabricate a “people“—a collective emotional state based on indignation, longing or fear—than to justify the inequality that is gnawing away at democracy. The result is a “post-truth-democracy” (Habermas 2022, 51). Once our political representatives and even democratic institutions, such as the judiciary, have been delegitimized, the way is open for the ongoing rise of *autocracies* and their unfortunate consequences. The appeal to competence and expertise, which is no longer identified with or expected from political representatives, but from their professional career or their ability to “recover” past glories and manage non-existent dangers.

Autocrats can learn to use elections to their advantage in a competitive authoritarian regime. Thus, elections and universal suffrage—even the division of powers—are no longer the privileged guarantors of legitimacy. Candidates tell

everyone what they want to hear. There is even talk of the decoupling of politics from truth as a direct effect of the impact of technology on democracy. If reality is a construct, whoever possesses the technology to convert facts into data has the power to undertake the formation of both an individual and a collective will —and this is a power that is a long way from deserving to be called democratic.

Bibliography

Apel, K.O. 2000. Globalization and the Need for Universal Ethics. *European Journal of Social Theory* 3 (2): 137–155.

Borges, Jorge Luis. 1960. Del rigor de la ciencia. In *El hacedor*, ed. J.L. Borges. Buenos Aires: Emecé Editores.

Conill, J. 2019. *Intimidad corporal y persona humana*. Madrid: Tecnos.

Cortina, A. 1993. *Ética aplicada y democracia radical*. Madrid: Tecnos.

d'Ancona, M. 2017. *Post-truth: The New War on Truth and How to Fight Back*. Random House.

Dahlberg, L. 2011. *Re-constructing Digital Democracy: An Outline of Four 'Positions'*. New Media & Society, September 2011, 13. Dunagan, J.F. *Politics for the Neurocentric Age*. op.cit., p. 59.

Dahlberg, L., and E. Siapera, eds. 2007. *Radical Democracy and the Internet: Interrogating Theory and Practice*. New York: Palgrave.

Dunagan, J.F. 2004. Neuro-Futures: The Brain, Politics and Power. *Journal of Futures Studies* 9 (2, November): 1–18.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Finn, E. 2018. *What Algorithms Want: Imagination in the Age of Computing*. Cambridge: MIT Press.

Fuchs, C. 2017. From Digital Positivism and Administrative Big Data Analytics Towards Critical Digital and Social Media Research. *European Journal of Communication* 32 (1): 37–49.

García-Marzá, D. 2005. Trust and Dialogue: Theoretical Approaches to Ethics Auditing. *Journal of Business Ethics* 57 (3): 209–219.

———. 2013. Neuropolitics and Democracy: A Necessary Dialogue. *Daimon International Review of Philosophy* 59: 171–182.

———. 2021. Posverdad. In *Glosario de patologías sociales*, ed. M.S. Cenci and J. Pizzi. Brasil: UPFEL.

Generalitat Valenciana. 2018. *Estrategia de Inteligencia Artificial de la Comunidad Valenciana*.

Gillespie, T. 2012. Can an Algorithm Be Wrong? *Limnologica* 1 (2).

———. 2016. # trendingistrending: When Algorithms Become Culture. In *Algorithmic Cultures*, 64–87. Routledge.

Habermas, J. 1971. *Toward a Rational Society: Student Protest, Science, and Politics*. Vol. 404. Boston: Beacon Press.

———. 1987. *Knowledge and Human Interests*. Boston: Polity Press.

———. 2018. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Wiley.

———. 2021. Überlegungen und Hypothesen zu einem erneuten Strukturwandel der politischen Öffentlichkeit. In *Ein neuer Strukturwandel der Öffentlichkeit?* 470–500. Nomos Verlagsgesellschaft mbH & Co. KG.

———. 2022. *Ein neuer Strukturwandel der Öffentlichkeit und die deliberative Politik*. Frankfurt: Suhrkamp Verlag.

Harari, Y.N. 2016. *Homo Deus: A Brief History of Tomorrow*. Random House.

Harsin, J. 2015. Regimes of Posttruth, Postpolitics, and Attention Economies. *Communication, Culture & Critique* 8 (2): 327–333.

Hill, R.K. 2016. What an Algorithm Is. *Philosophy and Technology* 29 (1): 35–59.

Horkheimer, M. 1972. *Critical Theory: Selected Essays*. Vol. 1. A&C Black.

Husserl, E. 1970. *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*. Northwestern University Press.

Jones, M.L. 2018. How We Became Instrumentalists (Again) Data Positivism Since World War II. *Historical Studies in the Natural Sciences* 48 (5): 673–684.

Kant, I. Akademische *Kant's gesammelte Schriften*. Royal Prussian Academy of Sciences.

Kitchin, R. 2017. Thinking Critically About and Researching Algorithms. *Information, Communication & Society* 20 (1): 14–29.

Lakoff, G. 2008. *The Political Mind*. New York: Viking.

———. 2014. *The All New Don't Think of an Elephant: Know Your Values and Frame the Debate*. Chelsea Green Publishing.

Larson, E.J. 2021. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Cambridge: Harvard University Press.

Lash, Scott. 2007. Power After Hegemony: Cultural Studies in Mutation? *Theory, Culture & Society* 24 (3): 55–78. https://doi.org/10.1177/0263276407075956.
[Crossref]

Lukes, E. 1985. *El poder, Un enfoque radical*, XXI. Madrid: Siglo.

Lukes, S. 2021. *Power: A Radical View*. Bloomsbury Publishing.

Mager, A. 2012. Algorithmic Ideology: How Capitalist Society Shapes Search Engines. *Information, Communication & Society* 15 (5): 769–787.

Mair, P. 2013. *Ruling the Void: The Hollowing of Western democracy*. Verso Books.

Mittelstadt, B.D., and L. Floridi. 2016. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. In *The Ethics of Biomedical Big Data*, 445–480. Springer.

Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3 (2).

Morozov, E. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Neumann, G., G.E. Marcus, G. Grigler, and M. Mackuen. 2007. *The Affect Effect- Dynamics of Emotions in Political Thinking and Behavior*. Chicago: University of Chicago Press.

O'neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway books.

Pariser, E. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. New York: Penguin.

Rehg, W. 2015. Discourse Ethics for Computer Ethics: A Heuristic for Engaged Dialogic Reflection. *Ethics and Information Technology* 17 (1): 27–39.

Striphas, T. 2015. Algorithmic Culture. *European Journal of Cultural Studies* 18 (4–5): 395–412.

Taddeo, M., and L. Floridi. 2018. How AI Can Be a Force for Good. *Science* 361 (6404): 751–752.

Tsamados, A., N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo, and L. Floridi. 2022. The Ethics of Algorithms: Key Problems and Solutions. *AI & Society* 37 (1): 215–230.

van Dijck, Joseph. 2016. *The Culture of Connectivity: A Critical History of Social Media*. New York: Oxford University Press.

Weber, M. 2019. *Economy and Society: A New Translation*. Harvard: Harvard University Press.

Westen, D. 2007. *The Political Brain*. New York: Public Affairs.

Wolkenstein, A., R.J. Jox, and O. Friedrich. 2018. Brain-Computer Interfaces: Lessons to Be Learned from the Ethics of Algorithms. *Cambridge Quarterly of Healthcare Ethics* 27 (4): 635–646.

Wolton, D. 2018. Information and Communication: An Urgent Theoretical Situation. *Hermes, La Revue* 82 (3): 10–18.

Wrong, D.H. 1997. *Power*. London: Transaction Publishers.

Yuste, R., S. Goering, G. Bi, J.M. Carmena, A. Carter, J.J. Fins, et al. 2017. Four Ethical Priorities for Neurotechnologies and AI. *Nature* 551 (7679): 159–163.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_10

10. Dialogic Digital Ethics: From Explicability to Participation

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

> *AI technologies can be of great service to humanity and all countries can benefit from them, but they also raise fundamental ethical concerns, for instance regarding the biases they can embed and exacerbate, potentially resulting in discrimination, inequality, digital divides, exclusion and a threat to cultural, social and biological diversity and social or economic divides; the need for transparency and understandability of the workings of algorithms and the data with which they have been trained; and their potential impact on, including but not limited to, human dignity, human rights and fundamental freedoms, gender equality, democracy, social, economic, political and cultural processes, scientific and engineering practices, animal welfare, and the environment and ecosystems (…). (UNESCO, Recommendation on the Ethics of Artificial Intelligence 2021, 3)*

Abstract
The aim of this chapter is to set out a proposal for dialogic digital ethics based on a close reading of the "Ethics Guidelines for Trustworthy AI" written by the European Commission's High-Level Expert Group on Artificial Intelligence. These need to be digital ethics whose normative horizon of action—their criterion of justice—lies in dialogue and the possible agreement of all the agents involved in and affected by the digital reality. In line with this criterion, it is proposed that such ethics be inserted into the democratic system as a structural element of civil society. The aim is to show that the participation of all parties involved is not only advisable, but morally required for their endeavour to generate a common European will and governance in response to the current

fourth Industrial Revolution. The recognition of universal dignity as implied by person-centred Artificial intelligence is unthinkable without the possibility of equal participation, both within the state and civil society. Without participation, trust cannot be built or guaranteed. As we intend to show, the principle of explicability plays a decisive role in achieving this goal as a principle that has a moral, and not merely an instrumental, value.

Keywords Artificial intelligence – Digital ethic – Dialogical ethics – Principle of explicability – Institutional design – Trustworthy

10.1 Towards Dialogic Digital Ethics

While our leaders insist and repeat that all technological progress must be accompanied by an ethical vision, a humanist basis, a project for progress, innovative and inclusive development, and so on, political strategies, public policies, major technological projects, public-private contracts and political decision-making contradict this. Such language comes from marketing, with a corresponding mix of desire, reality and exaggeration. There is considerable investment behind the mere announcement of what is technologically possible.

If the biases of superiority, neutrality and objectivism that consistently accompany the concept of digital transformation are not denounced; if these prejudices are not revealed; digital ethics—the analysis of what is right or wrong in the technological and digital sphere—can only be applied after the problem has appeared, in the wake of the consequences of the decisions and actions already taken on the basis of this technological rationality being reduced to data and Big Data, algorithms and Artificial intelligence. As experience shows us on a day-to-day basis, this will already be too late (Apel and Apel 1988; Rehg 2015; Yuste et al. 2017).

The refusal to recognize any human intervention, as if data were “out there” waiting to be discovered and the algorithms were not *de facto* the result of our interpretation of reality and our interest in knowledge, not only demonstrates a clear inability to understand the situation, but also a clear intention to *reify,* and therefore *obstruct* any possibility of ethical design and governance (García-Marzá 2005). Making this accusation is the first task of dialogic digital ethics, but not the only one, as we will see below.

More than 20 years ago, the European Commission, in its famous Green Paper, *Promoting a European Framework for Corporate Social Responsibility* (COM 2002, 3), which stated that social responsibility would be a key issue for the European economy and become, within 10 years, “the most competitive and

dynamic knowledge-based economy in the world capable of sustainable economic growth with more and better jobs and greater social cohesion". However, the result could not have been more disappointing. Today, a concept like social responsibility not only fails to generate trust, but has become a merely superficial concept or, even worse, a concept used to conceal clearly unjust situations (greenwashing). In view of the factual impossibility of avoiding its merely instrumental and manipulative use, some think it would be best to abandon the concept altogether (García-Marzá 2020). We know all too well that voluntary commitments do not usually go very far when confronted by the bottom line of financial results. Can the same be true of the European Commission's current commitment to its strategy of positioning ethics as a key factor in strengthening citizens'trust in digital development, while at the same time generating a competitive advantage for European AI companies? Our thesis consists of affirming that on this occasion the introduction of the principle of explicability brings a key differentiating element—*the participation of all those affected*—which can at least obstruct, if not prevent, this merely instrumental use of ethics and, at the same time, foster ethical innovations.

Digital ethics, concerned with the moral foundations underlying the trust we place in different digital practices and their respective technologies, must begin by criticizing the prejudices that once more reappear linked to the technological domain presupposed by science, whether as scientism, objectivism or positivism, and today dressed up with neuroscience and AI. The starting point for digital ethics can only be to overcome this obstacle: the epistemological dogmatism that prevents us from dealing with the human intervention that defines and manages all kinds of technologies. If an algorithmic decision is more just than a human decision, our moral and political autonomy disappears. We must show once again and tirelessly insist that the definition, identification and selection of data always depend on specific interests, and when this are used to construct algorithms, they are, and can only be, value-laden (Mittelstad et al. 2016; Floridi 2019).

Hence, the European Commission, in its 2019 report "Building Trust in Human-Centric Artificial Intelligence", states that: "AI brings new challenges because it enables machines to "learn" and to take and implement decisions without human intervention. (...) Yet, decisions taken by algorithms could result from data that is incomplete and therefore not reliable, they may be tampered with by cyber-attackers, or they may be biased or simply mistaken" (COM 2019:2). Like all other technologies, digital technologies are a tool: a means to achieve certain ends, not an end in themselves. As we shall see, the meaning of a piece of data—why one piece of data is significant and another is not—is always

a human decision. Algorithms depend on a reality built upon a given interest. An algorithm that gathers, integrates and makes sense of the data is "fabricated" on the basis of this interest. Networks and the Internet of Things are the source of Big Data, and algorithms sort this data. Finally, AI provides us with the Artificial intelligence systems and their technologies required to obtain the desired results.

This critical perspective does not operate outside reality. In November 2021, UNESCO, held its 40-first General Conference, presented the project *Recommendations on the Ethics of Artificial Intelligence*, which identified the problems we have been highlighting, and denounced this ideology of neutrality while warning of the dangers it entails (UNESCO, 2021). Meanwhile, European documents had already begun highlighting the importance of ethics for trustworthy Artificial intelligence.

"Opinion 4/2015 of the European Data Protection Supervisor" entitled "Towards a new digital ethics. Data, dignity and technology" sought to stimulate an open and informed debate on the definition, justification and application of *new digital ethics* (2015: 5). A debate involving civil society, designers, business, academics, public authorities and regulatory authorities and ethics that seek to "realize better the benefits of technology for society and the economy in ways that reinforce the rights and freedoms of individuals". The doubt remains the same: are we facing a new effort to mask the injustices caused by the new digital technologies and the big technology companies, who are today the masters of globalization, albeit in the guise of working to protect human dignity?

As applied ethics, digital ethics aim to explicitly set out and manage the ethical foundations of trust in the digital sphere across all its applications, practices and institutions. They are conceived as applied ethics because they are not intended to justify the moral knowledge we use throughout the process of digitization, but rather to apply this knowledge in the field of professional ethics, in institutional design and in the development and monitoring of public policies. The primary good provided by digital practice is precisely data as meaning attributed to signs, as interpretations of what is encountered, of reality. In fact, a single item of data is a formal reduction of what is encountered in the real world. The data does not represent reality—the given—but what we are interested in, and it is not collected on the basis of a logic of cause and effect, but rather a logic of correlation, trends and patterns. Neutrality remains outside this logic, as that is a choice to define a given fact as a given piece of data and then to assert that this is, for example, the essence of human behaviour. Floridi and Taddeo defines the main tasks of digital ethics as follows:

> (…) a new branch of ethics that studies and evaluates moral problems

> related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including Artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values) (Floridi and Taddeo 2016, 1).

We can also talk about the ethics of Artificial intelligence, algorithmic ethics, data ethics, cyberethics, and so on, but in all these cases the concern and interest are the same: to give an account of the significance and value of the ethical dimension today in global contexts without a global legal order. It must not be forgotten that this dimension is characterized by its claim to universality. In fact, today, digital ethics, with its values, principles, guidelines, and so on, are already involved in the world of technology much more than any other force. They act as the perception and assessment of what is morally good, right or just, and influence public opinion, what is socially acceptable or preferable, and what is politically feasible, and therefore, ultimately, what is legally enforceable. Unfortunately, this power to intervene is, as a rule, only deployed when negative consequences arise or are expected, i.e., when the damage has been done.

Digital ethics are understood as applied ethics whose scope of action is digital practice, its processes and technologies; its decision-making systems and, of course, the institutional frameworks in which they are produced. They also concern the ethical impacts and implications of digital technologies, the problems they present and the digital reality they construct, whose primary asset is data and the technological possibilities opened up by the use of Big Data. Floridi refers to the *infosphere* as a set of data-driven and data-dependent practices; digital ethics are concerned with the scope and rules that enable interactions in this new digital sphere. The problem is no longer digital innovation but rather its governance (Calvo 2021).

10.2 From the Desirable to the Possible

As is the case for other applied ethics, the generation and maintenance of trust in the digital sphere require three basic steps stemming from the impossibility of automatically putting ethical principles into practice, as they continue to be abstract moral obligations. We therefore need to distinguish three levels along the path from moral obligation to its practical realization: the level of justification focused on the principles' moral validity; the level of institutional

adequacy, centred on the design of possible governance; and, finally, the level of organizational precision concerning the application of each concrete and specific practice (García-Marzá 2017).

The idea that it is not possible to move directly from ethical principles to digital practices in need of regulation can also be found in the European Commission's proposal from the High-Level Expert Group on Artificial Intelligence "Ethical Guidelines for Trustworthy AI" (2019). This proposal leaves no doubt about the role expected of ethics: "Our vision is to ensure and scale Trustworthy AI" (par.13). Experience warns us of the risk that these guidelines will remain just mere words, and that they will not be applied with due value and effectiveness to a digital revolution that has already been underway for some time. Much is said, but little or nothing is done. The result is not trust but, on the contrary, suspicion and mistrust. Our discussion above concerning trust should be remembered here.

We speak of *trustworthiness* when we believe we have reason to expect something and when we think that this or that expectation will be fulfilled. It refers to the quality that a person or company is trustworthy. *Reliability*, on the other hand, refers to the probability that something works well. The willingness of one party to depend on another is not accidental, nor is it based on arbitrary intuitions removed from experience. To trust is not based on a feeling or irrational belief; on the contrary, it is reasons that make us trust, give us courage, drive and encouragement to act, and motivate us to action (García-Marzá 2004). The European Strategy and its Coordinated Plan make it clear that "Trust is essential to facilitate the uptake of AI technologies" (European Commission 2021, 31). Artificial intelligence, the Commission reminds us, is not an end in itself, but a means to "improve""people's lives. These are the expectations we have and upon which digital practice rests its legitimacy and social credibility.

The normative framework developed in this European Commission document also offers three different levels for making the transition from ethical principles to reality; for achieving what "ought to be". As we shall see, none of these three levels can be ignored, yet we must also be aware of the normativity that accompanies each of them. These three levels are the same ones that characterize all applied ethics and constitute its general framework for action:

1. The *foundations* of trustworthy AI are to be found in the Declaration of Human Rights and the European Charter of Human Rights. These are ethical principles based on fundamental rights and, above all, on the recognition of the equal dignity of all people. These are: *autonomy* (respect for human

dignity and recognition of the other as a valid interlocutor); *non-maleficence (*prevention of harm and protection of dignity*)*; *justice* (inclusion and fair distribution of costs and benefits) and *explicability* (transparency, information and participation in decision-making). The aim of these ethical principles is to inspire and guide the rationale for the development, use and application of AI systems and thereby define their responsibility—that for which they are accountable to society.

2. The *achievement* of trustworthy AI is based on seven basic requirements that define the conditions for its implementation throughout the life cycle of AI systems: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; social and environmental well-being; and accountability.

3. Its specific *implementation*, in turn requires an analysis of the conditions and characteristics of the specific contexts; an *assessment of* the real possibilities for action.

It is beyond the scope of this article to address the intricacies of this architecture. However, we can affirm that we are dealing with a solid regulatory framework in which, for the first time ever in European Union guidelines, ethics play a central role in both defining and governing human-centred Artificial intelligence. The European Commission is aware that the trustworthiness sought depends upon both legal compliance and robust technological conditions. Yet it also, and especially, depends on the voluntary and informed consent of all the agents and processes that form part of the socio-technical context; on the consensus of all the individual and institutional agents that participate in generating and managing this *infosphere*. The guidelines state: “In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system’s life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment” (para. 79, 18).

This requirement derives directly from the principle of autonomy as a recognition of the equal dignity of all people. The core support for generating trust in a dialogic digital ethic is precisely this free and equal participation; this concern for deliberation and search for agreement. Authors such as K.O. Apel and J. Habermas have developed the foundations of discursive ethics, and, more recently, a second wave of thinkers have addressed an applied ethics derived from this moral requirement for participation as a requisite for the development

of autonomy, and as a support for the intrinsic value of dignity (Cortina et al. 2008). The moral basis for the need for this dialogue and possible consensus is to be found in the recognition of the intrinsic value of the dignity of all people involved in the digital reality (para. 41, 10).

The moral value of the guidelines under analysis, and with it their binding force; their obligation and their capacity to become moral resources available to all parties involved; do not disappear at the second and third levels. Instead, at these levels of adequacy and specificity, the ethical principles of the first level must be integrated with the possibilities for realization, as well as the limits and potentialities that reality offers us. Morality, and therefore the enforceability of decisions, actions or institutions, does not vanish following the application of these principles, and this is demonstrated by our capacity to morally assess the results achieved, upon which, as we will see in the following section, the ethical foundations of trust are based.

However, the concern for *moral traceability* is not clearly discerned in the European Commission document. It is only when we enter the realm of requirements or conditions for the application of the principles that we encounter the division of human agency and oversight, which are segregated into three levels:

1. *Human participation*: the ability to involve human beings in all decision cycles of the system, which in many cases is neither possible nor desirable.
2. *Human control*: the ability to involve humans during the system design cycle and when monitoring the system's performance.
3. *Human oversight*: the ability to monitor the overall activity of an AI system, including its economic, social, legal and ethical effects, as well as the ability to decide when and how to use the system in a given situation (High-Level Expert Group on IA 2019, 16)

When it is claimed that the intervention of all human beings is "*neither possible nor desirable*", the moral validity of participation disappears. As a result, the whole system of a reliable Artificial intelligence loses its moral value and thus its strength and effectiveness. Participation ranges from being seen as a condition of trust, and from being morally required, to it being solely "advisable" to consult stakeholders who may be directly or indirectly affected by the system. Moreover, if the participation of all those affected is "undesirable", how can we trust a system that does not depend on us? How can we know the

effects of a technology on individuals and groups, for example, on the autonomy and freedom necessary for a democratic society, when ethical principles are no longer required but recommended? How can we manage, or even know, the limits of technology? How can we define what is possible and what is impossible? And how can we halt what is undesirable, right at the design stage, if we can only agree or refuse what we barely understand?

The transition from the "*everyone*" reflected in ethical principles (level 1) to the *"everyone"* possible in different practices and institutions (levels 2 and 3) is one of the most important dimensions of ethical reflection and the key to turning moral knowledge into a moral resource: the ultimate goal of all applied ethics (García-Marzá 2004). Only in this way can we address the ethical basis of trust in Artificial intelligence. We are not saying that all those potentially affected can participate, but that anyone can participate. In my view, it is this step that provides us with the principle of explicability. The participation of all those affected is not only advisable, it is morally required. Another question is how to do so in complex, opaque and global contexts such as those making up today's digital reality.

10.3 The Principle of Explicability: From Transparency to the Need for Participation

The principle of explicability can answer many of the questions posed above as long as it does not lose its moral character, i.e., is not reduced to being an alternative to the universalism and inclusiveness stipulated from a moral point of view. The reason for adding a new principle to those already used by bioethics is not only the complexity of Big Data, the Internet of Things (IoT) and algorithms, nor the disproportion between those who build the algorithms and the decision-making and learning machines, as well as those who will suffer the consequences. This asymmetry not only gives rise to a need to understand and account for AI decision-making processes (Floridi et al. 2018, 699), but also to reverse the lack of reciprocity and mutual recognition through free and equal participation, the possibility for effective influence, the discussion of possibilities for action, and the limitation of technological developments that remain unclear with regard to the consequences according to the well-known principle of participation. Such a need is not merely strategic, nor does it derive only from the complexity, opacity and unintelligibility of digital practices, but also from the morality of a principle that allows us to move from the desirable to the possible, as it is capable of internally linking principles, requirements and

contexts.

In the document under discussion, the principle of explicability is presented as “crucial” so that can users trust AI systems and maintain this trust. This means:

> (...) that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions—to the extent possible—explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested (High-Level Expert Group on IA 2019, 13).

Across the range of current international guidelines, this principle is expressed in different terms, with greater or lesser semantic identity: transparency, accountability, intelligibility, comprehensibility, communicability, openness, clarity, etc. (Jobin et al. 2019; Mittelstad et al. 2016). It is a principle that encompasses the knowledge of how an algorithm works and also who is responsible for its operation. Thereby, this new principle becomes the thread that establishes a binding coherence for the rest of the principles, as it is a matter of the citizen knowing and understanding the digital reality in which he or she is immersed. According to Floridi et al.,

> For AI to be beneficent and non-maleficent, we must be able to understand the good or harm it is actually doing to society, and in which ways; for AI to promote and not constrain human autonomy, our “decision about who should decide” must be informed by knowledge of how AI would act instead of us; and for AI to be just, we must ensure that the technology—or, more accurately, the people and organisations developing and deploying it—are held accountable in the event of a negative outcome, which would require in turn some understanding of why this outcome arose. More broadly, we must negotiate the terms of the relationship between ourselves and this transformative technology, on grounds that are readily understandable to the proverbial person “on the street” (Floridi et al. 2018, 790).

However, things are not that simple. The definition of the principle of explicability in the *Guidelines* neither implies, nor requires the participation of all those affected, upon which, it should be recalled, its moral value is based. Instead, it seems more akin to a purely informative and therefore vertical process. We read about the transparency, explicability and accountability of

decisions taken, about who has decided to act and how, and so on. These features are as absolutely necessary as they are completely insufficient: they do not require the presence and agreement of all the parties involved. We may think that little is better than nothing, but once more what is being questioned is the moral value of the principle of explicability and, thus, its force. This aspect becomes apparent just by reading the guidelines document further: "An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g., traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights" (High-Level Expert Group on IA 2019, 13).

Seemingly, as is argued, "the degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate" (High-Level Expert Group on IA 2019, 13). While we agree with this, does it only apply in these cases? Who decides on any possible consequences at the time of data collection? Who defines the data to be fed into the system? As has been clearly established, autonomy depends on the possibility of participating in everything that affects us. Our dignity and the moral value of this principle are based upon this capacity. But the moral value disappears if its universality is not taken into account at all. According to the above text, the degree of need for explicability depends to a large extent on the context and severity of the consequences of an erroneous or inappropriate outcome. Who decides whether the context is appropriate or not? Who decides the severity of the consequences? Without an answer, the three ethical principles preceding the principle of explicability easily lapse into a "justify-it-all".

In our view, traceability and auditability, as well as clarity and intelligibility, are indeed part of any transparency process. However, transparency is only half of explicability. Participation is needed, as it enables us to deliberate on the questions we have been leaving unanswered. For example, it enables us to reach agreements on whether or not we should have to accept certain disadvantages. In the absence of free and voluntary acceptance, there is no autonomy. In turn, without autonomy there is no moral basis to support the recognition of the equal dignity of persons. It is this that is the moral basis of the principle of explicability.

At first glance, it may seem naïve to demand the inclusion of *all* possible stakeholders in decisions that affect them when we are talking about billions of

people, and complex and often incomprehensible processes, as well as consequences that no one expected or foresaw, and so on. Yet such candour disappears when we recall the three steps necessary between moral demand and reality; from the duty “to be” of ethical principles to the specific situation and reality of the socio-technical, social and economic contexts.

We must therefore show that the moral validity of the ethical principles presented, the requirement of universality and thus their conviction value and force, continues to be present when we descend to the level of realization—the necessary requirements for enabling and guaranteeing the application of the principles while always being mediated through the empirical conditions of realization. This is, in our opinion, the key role played by the principle of explicability, as it allows us to speak of a *moral strategy*: a better or worse approximation to the idea of the participation and the possible agreement of all parties involved. In our opinion, a comparison of the principle of explicability with the *principle of publicity* introduced by Kant in *Perpetual Peace: A Philosophical Sketch* allows us to understand how digital ethics operates. As applied ethics, these must give a reason for the “ought to be” embedded and operative in the reality of digital practice and in the construction of the trust they deserve.

10.4 Institutional Design: The Role of Applied Ethics in the Democratic System

Kant proposed the *principle of publicity* as a means of mediating between ethical principles and practice —between the desirable and the possible— yet without this mediation losing its moral value (García-Marzá 2012). A review of some of the issues raised by his proposal helps clarify the moral value of the principle of explicability along with the demand for participation attached to it. When Kant warns us of the dangers of institutions in which *everyone, without being everyone, decides*, he warns us of the consequences of annulling the difference between the moral requirement, the second *everyone* in the sentence, and its practical realization, the first *everyone*. This distance between ethical principles and their possible practical realization can never be completely bridged, irrespective of how well justified the participatory structure of the company, the university, the parliament, the laboratory, etc. may be. It is precisely within the greater or lesser distance between the “moral everyone” and the *factual everyone* where any critical perspective is based and it is from there that the arguments supporting trust are constructed.

This is the basic idea that Kant develops through the principle of publicity, and its first definition speaks of the injustice or immorality of the decisions taken, simply stating: “All actions relating to the right of other men are unjust if their maxim is not consistent with publicity” (ZeF, VIII, 381). Applied to our field, if algorithms have to be secret, it is clearly and emphatically because they are unfair. We can propose property rights, patents, and so on, but moral validity is inversely proportional to opacity. If our decision or action—the premises, for example of the origin of the data—cannot be made public, it is because they are unfair. According to this principle, any element of Artificial intelligence systems that can neither be communicated or explained is immoral. This means that it should be discarded. Secret algorithms cannot be ethical.

However, Kant also offers a positive definition of the principle of publicity, albeit one that is not as clear-cut as the negative one. He thus states: “All maxims which stand in need of publicity in order not to fail their end, agree with politics and right combined” (ZeF, VIII, 386).

According to this second principle not only do reasons, decisions, procedures, and so on have to be announced and made known, but reference must also be made to a kind of union or concurrence of all those concerned. The possibility of public knowledge entails, so to speak, the need for the acceptance or approval of others, otherwise it could fail to achieve its aims. Essentially, on the basis of its consent, public now takes on a second meaning: it no longer only opposes secrecy, it also seems to oppose what is closed, as it requires the approval of the public in order to be fulfilled (Chambers 2002). The need for publicity means that algorithms, and other elements of AI systems, could not be successful or effective without being public. This is simply due to the distrust they generate.

Needless to say, a possible agreement cannot be anticipated by one or more of the parties involved, for example programmers, companies, government agencies, and so on. While we cannot anticipate agreement, there are mechanisms for knowing whether we are moving away from or towards the idea of a possible consensus of all those involved. At the end of his book, Kant proposes a further step in this relationship between principles and their realization, with an explanation of the critical potential of the principle of publicity (García-Marzá 2012). As if providing a third formulation, Kant equates publicity with the elimination of all distrust. He states: “(…) if a goal can only be achieved with the help of publicity, it means that there is no distrust in the underlying political maxims that are congruent with the goals and rights of all” (ZeF, VIII, 386). Is this not the trust that ethical guidelines seek?

Therefore, the moral, and not only the instrumental, justification of the

principle of explicability derives from the recognition and respect for the autonomy of all those affected by or involved in regulation or legislation. Kant refers this justification to public reason, which he defines as that faculty *where everyone has voice*. The moral justification for the principle of explicability is based on the involvement of all those involved and affected, both people and institutions. The three formulations are derived from the reciprocal recognition of all those involved in the institutional regulation of the irreplaceable character of free will, on which the dignity of the individual is based. It is this will that requires our consent or agreement: our free acceptance. It is only on the basis of this conformity that it is possible to build and guarantee trust. By using this moral justification, it must *always* be possible to apply the principle of explicability, understood as the sum of transparency and participation.

Without this principle, which is capable of mediating between the moral *everyone* and the pragmatic *everyone*, every other principle becomes meaningless and we are left without a criterion for right and wrong. Yet this criterion must be applied, as is discussed below, from the outset, by defining the digital contribution sought from the origin of the data and through the construction of the algorithms. This double dimension of the principle of explicability includes *responsibility* and the capacity to present arguments to those affected for what we do, or do not do. Moreover, responsibility always becomes co-responsibility, giving rise to the generation of trust.

The seven requirements defining the second level of the implementation process set out in the document do not deny this need for participation, but it is discussed only as a possible undertaking, as part of a horizon of action that *guides, but does not oblige*. Instead of this renunciation of moral enforceability, we should consider an *institutional design* capable of creating spaces where this free and equal participation would be possible.

Clearly trust has a psychological component, as it does not cease to be a belief in a person or organization's future behaviour. However, it also has a rational basis, a set of intersubjective arguments that underpin this belief and motivate our perception and assessment as developers, researchers, managers, employees, and so on. Among these arguments, the knowledge that we are dealing with organizations that are committed to behaving responsibly and can demonstrate their commitment to ethical guidelines occupy a privileged position. If trust is based on having reasons to expect something, this information can motivate our decision and subsequent action (García-Marzá 2004).

With regard to this motivation, and in order to build trust, a declaration of good intentions from professionals or their organizations is insufficient. Drawing on the principle of explicability as an ethical principle, any information

management that aims to account for compliance with the "Ethical Guidelines for Trustworthy AI" must, for each specific scenario, engage in making public the efforts made. It is not just a question of a willingness to be honest, but of this willingness acquiring the status of a *public commitment*, in the two-fold sense of transparency and participation, as has been discussed. Institutional design theories work with this idea by referring the capacity to produce trust to this "discursive justification potential" (Goodin 1998).

Designing seems a pretentious and risky term, but this first impression disappears when we realize that its etymological root *designare* indicates the task of signalling which institutionalization of the requirements is the most appropriate for a particular context. While designing or redesigning are intentional activities, they should always be understood as contributions to public deliberation about which ethical infrastructure is most appropriate so that our organizations, and their internal and external communication, can build trust. In other words, the concern is how to sustain and develop an organization's credibility and reputation. In order to respond to this public justification, it is not enough to rely on a professional's goodwill; instead we must have processes and organizational structures in place that enable and enhance the ethical guidelines outlined above.

To anchor these ethical foundations of trust, transparent, traceable and intelligible information is required in conjunction with the possible participation of stakeholders or their representatives "by design". There is no transparency without the possibility of participation; without being able to decide, for example, what is reported or how the consequences are calculated and for whom. Or in other words, the calculation of the moral hazard—the possibility that others will suffer the consequences of any decisions or practices implemented, ones which we did not expect would happen—is directly proportional to participation. Put more simply, there is no autonomy without the possibility of being included in the decisions that will ultimately affect us. As the document rightly points out, this participation forms part of the whole life cycle of the algorithm, as well as its creation. We must not only accompany technology, we must stay ahead of it, always remaining one step ahead (Etzioni and Etzioni 2017; Dignum 2018).

This institutional design and ethical infrastructure should have the following building blocks, regardless of the size and complexity of the decisions made and the organizational spaces where they are taken: a Code of Ethics and Conduct (including good governance); a Sustainability Report or Social Responsibility Report; an Ethics Committee; an Ethics Line and an Ethics Audit. We discuss this infrastructure in greater detail in the following chapters, and we will discuss

how they constitute a proposal for the generation of *civil public spheres*, spaces for participation within the different institutions of civil society.

For now, it is not possible to discuss each element that forms part of the ethical infrastructure needed to convert the moral knowledge presupposed by digital intelligence and its technologies into moral resources (García-Marzá 2004). However, we must underscore the fact that for many organizations, codes of ethics are all they have communicating social responsibility, as if, in order to be credible, it was enough to declare goodwill to be credible. They forget that ethical codes are the just first step in building trust, not the last, and far from the only one. Hence, for this proposed ethical infrastructure in public and private institutions operating in the socio-technical context of digital practice, it is necessary for the compliance systems to be adopted to be implemented from the very outset. We refer to these codes of ethics as forming a new generation, as they have integrated the compliance mechanisms—the ethics committee and the ethics line—as the two core mechanisms for the participation of all stakeholders and not only management or ownership, as is usually the case. We have also referred to this as a renewal, indeed, as a new generation of codes of ethics, precisely because they incorporate participation not only in the form of a definition of commitments, but also through the monitoring and control of their fulfilment (García-Marzá 2017). This new generation of codes of ethics goes beyond current ethics and compliance programmes, which are always limited to the legal sphere and concerned only with criminal risk. We will return to this ethical infrastructure with regard to major technology corporations.

The English word stakeholders is sometimes translated into Spanish as “grupos de interés” (interest groups) fails to reflect the fact that individuals and groups can be involved and affected by something without being interested, or even aware of it (García-Marzá 2005; Parisier 2017). Such is the case in most cases of data mining and the formation of Big Data. This is further complicated by the opacity, globality and complexity of the socio-technical systems that make up digital practice. If we ignore the value of what is handed over to algorithms it is impossible to speak of autonomy, because every form of interest requires knowledge and approval. Together they form the basic contribution of the principle of explicability and constitute its moral value, which is derived from being the condition of our autonomy’s possibility. In other words, without this principle and its consequent moral requirements, there is no possibility of dialogue and agreement between all the parties involved in and affected by digital practice, including the environment and society at large.

Participation should include both *internal* groups (researchers, developers, managers, workers, etc.) and *external* groups (competing companies,

government agencies, consumers, civil society organizations, and so on). It is about establishing an advanced ethical culture in which the principle of explicability plays a central role: any process, formula or mathematics that cannot be explained to all participants within these public spaces of deliberation and monitoring must be rejected. Naturally, this applies as long as we want to ensure a firm basis for building trust. To understand how applied ethics can help with democratic construction, it is necessary to enter into the institutional structure of a two-way democracy, particularly in an expanded and radical concept of civil society. And this is a dimension that the European guidelines do not consider at all.

Bibliography

Apel, K.O.. 1988. *Diskurs und Verantwortung das Problem des Übergangs zur postkonventionellen Moral*. Frankfurt: Suhrkamp.

Calvo, P. 2021. El gobierno ético de los datos masivos. In *Tecnologías socialmente disruptivas. Dilemata. Revista internacional de éticas aplicadas,* ed. J.Ruedan, vol. 34, pp. 31–49.

Chambers, S. 2002. A Critical Theory of Civil Society. In *Alternative Conceptions of Civil Society*, ed. S. Chambers and W. Kymlicka. Princeton University Press.
[Crossref]

Comisión de las Comunidades Europeas. 2002. *Libro Verde Fomentar un Marc Europeo para la Responsabilidad Social de las Empresas,* Brussels, 18.7.2001 com(2001) 366 final.

Comisión Europea. 2019. Generar confianza en la inteligencia artificial centrada en el ser humano.

Cortina, A., J. Conill, and D. García-Marzá, eds. 2008. *Public Reason and Applied Ethics: The Ways of Practical Reason in a Pluralist Society*. Ashgate Publishing.

Dignum, V. 2018. Ethics in Artificial Intelligence: Introduction to the Special Issue. *Ethics and Information Technology* 20: 1–3.
[Crossref]

Etzioni, A., and O. Etzioni. 2017. Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21 (4): 403–418.
[Crossref]

European Commission. 2019. Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

———. 2021. COM(2021) 205 final. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Brussels. https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review.

European Data Protection Supervisor. 2015. Hacia una nueva ética digital, Dictamen 4/2015.

Floridi, L. 2019. Translating Principles Into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology* 32 (2): 185–193.
[Crossref]

Floridi, L., and M. Taddeo. 2016. What Is Data Ethics? *Philosophical Transactions of the Royal Society* A374: 20160360. https://doi.org/10.1098/rsta.2016.0360.
[Crossref]

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena. 2018. AI4People-an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (4): 689–707.
[Crossref]

García-Marzá, D. 2004. *Ética empresarial: del dialogo a la confiança*. Madrid: Trotta.

———. 2005. Trust and Dialogue: Theoretical Approaches to Ethics Auditing. *Journal of Business Ethics* 57 (3): 209–219.
[Crossref]

———. 2012. Kant's Principle of Publicity, Kant-Studien. *Philosophische Zeitschrift der Kant-Gesellsschaft* 103: 96–113.

———. 2017. From Ethical Codes to Ethical Auditing: An Ethical Infrastructure for Social Responsibility Communication. *El profesional de la información* 26 (2): 268–276.
[Crossref]

———. 2020. Contra el nuevo revisionismo democrático: el alcance del valor moral de la democracia en K.O. Apel. *Ethic@* 19 (3): 568–591.
[Crossref]

Goodin, R.E., ed. 1998. *The Theory of Institutional Design*. Cambridge University Press.

Jobin, A., M. Ienca, and E. Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1 (9): 389–399.
[Crossref]

Kant, I. 1795. *Zum ewigen Frieden. Ein philosophischer Entwurf*, (ZeF), AK, VIII.

Mittelstadt, B.D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 3 (2): 2053951716679679.
[Crossref]

Pariser, E. 2017. *El filtro burbuja: Cómo la web decide lo que leemos y lo que pensamos*. Taurus.

Rehg, W. 2015. Discourse Ethics for Computer Ethics: A Heuristic for Engaged Dialogic Reflection. *Ethics and Information Technology* 17 (1): 27–39.
[Crossref]

UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligenc*e. United Nations Educational, Scientific and Cultural Organization, https://unesdoc.unesco.org/ark:/48223/pf0000381133/PDF/

381133eng.pdf.multi.page=3.

Yuste, R., S. Goering, G. Bi, J.M. Carmena, A. Carter, J.J. Fins, et al. 2017. Four Ethical Priorities for Neurotechnologies and AI. *Nature News* 551 (7679): 159.
[Crossref]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_11

11. A Critical Civil Society: An Ethical Framework to Control the Power of Algorithms

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

> *First, civil society is no longer considered s system of needs understood primarily in economics terms. Now civil society is seen as a system of ideas, values, ideologies, and, yes, interests understood primarily in sociological and political terms. Second, the central role of hegemony leads to the semiautonomy of the words of ideas from the economy… Today, critical theory as one of the heirs to Gramsci and neo-Marxism sees civil society as a sphere of identity formation, social integration and cultural reproduction, and although economic relations and the state play a part in these functions, their roles are, or ought to be, supporting, not leading.*
>
> *Simone Chambers, "A critical theory of civil society" (2002, 91)*

Abstract
Given that any democratic proposal must account for the institutional skeleton that sustains it, the aim of this article is to show that this articulation cannot be limited to the state and its laws. Instead, it requires a vision of democracy based on a two-way perspective in which state and civil society complement each other, while being aware, as we have seen, that it is in civil society, and through the use of communicative action, that both political power and economic power are shaped and legitimized. It is within this context that this article proposes institutional ethics perspective for deliberative democracy based on *participation within institutions*. It aims to contribute two closely linked aspects to the discussion on the possibilities of achieving deliberative democracy: the value of

civil society and an ethical perspective on institutional design. Both aspects must be taken into account if we are to continue along the path of achieving a democratic theory capable of curbing and reinstating the claims made for algorithmic democracy. The possibility of establishing the meaning of a human-centred Artificial intelligence lies not only within the state, nor its laws, which are totally necessary but insufficient; instead it is located in the institutional structure of civil society.

Keywords Artificial intelligence – Critical hermeneutic – Deliberative democracy – Two-way democracy – Institutional design – Participation

11.1 Two-Way Democracy: The Value of Participation

Habermas has provided one of the most systematic and coherent contributions concerning the place and function of moral discourses in our democratic systems (Habermas 1992, 135; Vitale 2006). Although we have already discussed the Habermasian conception of deliberative politics, it is worth recalling that his methodology consists in the reconstruction of the intuitive knowledge, competences and problem-solving capacities available to us as participants in democratic processes. From a participant's perspective, theory does not have to describe or prescribe, but rather to make explicit and justify the assumptions of meaning citizens themselves ascribe to their participation. There is no contrast between the ideal and the real because it is a normative content which from the outset is inscribed, at least partially, in the social facticity of the observable political processes (Habermas 1992). This reconstructive methodology provides access to the knowledge that we, as citizens of a democracy, in fact use.

In *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, Habermas identifies this moral knowledge as the moral discourses that refer to the possible, free and equal agreement of all those involved and affected. These moral discourses are embedded in democratic reality itself and shape the critical perspective from which we assess the moral validity of ethical/political discourses, strategic discourses and, as a result of all these, legal discourses (Habermas 1992, 207).

It is from the point of view of *what is good for everyone*—that is, from the moral principle of universalization—that we can speak of a participatory or radical democracy, as it affects all areas requiring a common will and includes all those involved and affected (García-Marzá 2008). For Habermas, the

democratic rule of law constitutes the legal institutionalization of this network of discourses and negotiations, and of the conditions for communication and procedures that make it possible. Civil society is located within these conditions as the social basis of the public sphere.

In Habermas' view, our communicative freedom—our ability to deliberate, to transform and to reach agreements about the interests at stake—*cannot be delegated*, it demands free and equal participation. As such a requirement cannot be met in complex, plural societies, representation emerges on purely technical grounds, so its legitimacy always depends on an unreserved openness of institutionalized deliberations to the flow of information from disorganized public opinion (Habermas 1992). For the government to be democratic, merely being elected is not enough—it is a necessary but not sufficient condition. This is demonstrated by autocrats' use of elections in so-called competitive authoritarian regimes. There is no democracy without a strong civil society to give it credibility and legitimacy.

The outcome of this reconstruction is an understanding of the democratic process as the sum or complement of the *state*, understood as the legal institutionalization of the network of (always fallible and revisable) discourses, negotiations and compromises, and *civil society*: the informal processes that make up the social basis of public opinion. The latter is charged with monitoring the other discourses, as it constitutes, in Habermas' words, the voice of those potentially affected: that is to say it is the expression of generalizable interests (1992, 441). This conception of a *two-way democracy* foregrounds the value of civil society, along with its significance and importance, as an inseparable part of any conception of democracy (Habermas 1992, 369).

However, within this reconstruction, the role of moral knowledge—of the discourses centred on *how we should act*—is reduced to integrating it into legal discourses, as the latter compensate for the weaknesses of a will based "only" on practical reason. For Habermas, without law, our ability to act according to moral values and precepts has a narrow margin for action. He states that in complex societies morality can only be effective within a future moral sphere if it is transposed to the legal sphere (Habermas 1992). If we question the effectiveness of this moral knowledge, this always leads to it being complemented by the law and, at most, being a critical perspective within a rather ethereal concept of civil society as the basis for public opinion.

Nevertheless, it is clear that reality does not operate in this way, and that in the other spheres—economic, social, cultural, and so on—this moral knowledge also counts when it comes to explaining the mechanisms for coordinating action. It is not possible to give a reason for the legitimacy of the institutions that make

up these spheres without referring to justice as a claim to moral validity. For example, is the law sufficient to account for the social responsibility of large technology platforms? (Maschewski and Nosthoff 2021). Does it suffice to speak of external, legal regulation in a global order? Do the companies that design and develop algorithms have nothing to say about the justice or injustice of their applications?

Even if we accept that we cannot directly apply discursive ethics to democracy, the institutional mediation required to move towards deliberative democracy is not reduced to law. Ethics—the reflection on our capacity to act morally—has more to say about our democratic systems. While Habermas basically reduces the possibilities of applying moral resources to law, Cortina extends them to all areas or spheres of civil society, through her conception of the subject as a valid interlocutor and, therefore, capable of making decisions from the perspective of generalizable interests in all areas, not only in the political sphere. In applied ethics, it is not a matter of solving specific cases, the aim of casuistry, or the result of interdisciplinary work undertaken side by side with other approaches and methodologies, but rather of "designing the values, principles and procedures that should be taken into account by those concerned for each different case" (Cortina 1993, 168).

It is precisely this mediation of applied ethics that is one of the most innovative and important points for a revision of the relationship between ethics and democracy. Its importance is, once again, demanded by reality itself: "the dissatisfaction with the lack of participation in matters that affect us all has generated a multiplicity of ethical demands in social life, which are embodied in what has been called 'applied ethics'." These ethics share the demand encountered in the different fields where they applied: that it is those affected by the decisions who should not only have the last word, but also more than the last word (Cortina 1993, 176).

Whereas in Habermas this moral knowledge is only concerned with individual actions and decisions and must be translated into law when we speak of social agreements, applied ethics is presented as the development of our capacity to act according to our moral knowledge in the different spheres of civil society, be it in the ecological field or in technological development, the economic and business sphere, health and biology, the realm of information, and so on. In all these spheres, it is a matter of assuming the responsibility deriving from our autonomy—from our capacity to act according to our moral knowledge—and seeking agreements based on the reciprocal recognition of the equal dignity of all those involved and affected. This conception of the subject as a valid interlocutor is the best expression of this autonomy and it is accompanied,

as a condition of possibility, by participation. The definitive root of democracy is the autonomy of the individuals who make up a people, and the development of this autonomy is, in some way, accompanied by participation (Cortina et al. 2008).

From this point of view, participation acquires a moral value and not merely an instrumental and strategic one, because it is now about the conditions of possibility for the development of our autonomy, our communicative freedom and the opportunities to act accordingly. As the capacity to act according to our moral knowledge, this moral value of participation comes before deliberation and the search for agreements and refers directly to this idea of the subject as a valid interlocutor. Highlighting the value of participation, its moral value derives from its internal relationship with autonomy and, in our case, with the meaning of people-centred Artificial intelligence. Attending to people, for example, requires their participation in the three steps of any algorithmic construction: design, development and consequence calculation (Calvo 2018).

Although in this case, the term "democracy" is also seemingly reserved for the political sphere, the demand for participation extends to all spheres where decision-making affects us, and thus opens up the possibility of speaking of democracy beyond the strictly political realm. In other words, through this demand for participation, applied ethics become part of our understanding of democracy, as: "if we acknowledge that the ideal of *meaningful participation* claimed by participationists cannot be exercised in political life, because it would lead to an assembly model and other evils, it is necessary to respect other spheres in which this ideal can perhaps be embodied, while not permanently reducing individuals to the role of voters" (Cortina 1993, 16). It involves introducing critical hermeneutics in order to reconstruct the meaning and validity of the institutions of civil society, from companies to the media.

The so-called systemic approach to deliberative democracy, focusing on the possible realization of the deliberative ideal in our complex and technologically structured societies, has failed to analyse the participatory demands placed on this institutional structure of civil society, and thus no type of power other than political power is explained or justified. In other words, these types of power are misunderstood and have become unmanageable (Goodin 2008). A good example of this neglect—this disregard for the power of civil society—is provided by technology companies and research and development centres, and their failure to assume their responsibility for the problems we have discussed over the course of this book. Can we really say that the main actors, along with the state, who engaged in the construction of this algorithmic democracy, have nothing to do with the democratic system? Are they not, and do they never cease to be

authentic agents of justice or, in their case, injustice?

11.2 Applied Ethics: Structuring Civil Society

A participatory model of deliberative democracy that proposes to strengthen institutional structures in order to integrate the trends of algorithmic democracy into the democratic system—in other words, so that datafication and algorithms become no more than a means or instrument for achieving a more just society—is well aware of both the importance of legal regulation and a common legal framework capable of locating the new datafication technologies *within the democratic system*. Yet, it diverges from the sufficiency and effectiveness of all forms of external coercion, still more so in fast-paced, global and plural contexts. Legal remedies are, of course, necessary, but the question arises as to whether they are able to encompass both the coordination mechanisms present in the political sphere and those pertaining to civil society. In the latter case, we require a suitable methodology to approach the moral knowledge that underlies their practices and institutions, as well as a definition appropriate to their respective responsibilities. Let us not forget that democracy consists of a network of institutions, and those making up the state are just one part of it. Although the security provided by the coercive force of law stands in its favour, this requires credibility and trust (García-Marzá 2008).

As we know from Habermas, civil society may be understood as a network of communicative actions that constitute the social basis of public opinion, as it is within civil society that the network of content and positions that constitute public opinion is produced and reproduced. Their quality, in turn, depends on procedural criteria. This civil society is made up of associations, organizations and movements which arose in a more or less spontaneous way and whose nucleus institutionalizes the discourses that allow solutions to be found for problems concerning issues of general interest within the framework of more or less organized public spaces (Habermas 1992).

With regard to Habermas's formulation, we are solely concerned with highlighting the exclusion of the economic sphere: markets and companies; and the political-administrative sphere: parties, the judicial system and, in general, the entire parliamentary complex. As a result of this elimination, the possibilities for any intervention from civil society are reduced to the pursuit of generalizable interests that constitute and promote public opinion. In our view, if Habermas thereby succeeds in introducing the *burden of normative expectation*, and thus a critical perspective, into the theory of democracy, it seems to be at the price of sacrificing to the "systemic" realm the spheres of action in which the vast

majority of social relations are established, and where power is produced and reproduced. For Habermas, civil society can only function as a network of communicative actions. Solidarity is, therefore, the basic element of social coordination.

Civil society cannot, so to speak, be contaminated by the spheres where strategic action either prevails or interacts, as we have already seen in the first chapter (Honneth 1993). However, it must be highlighted that there is no system/lebenswelt dualism. Instead these two analytical dimensions are one and the same reality. There are neither associations nor organizations without strategic practices and power relations, nor institutions whose legitimacy does not depend on the possible agreement of those affected. Habermas seemingly overlooks those spheres of justice where communicative action is not exclusive and is unwilling to acknowledge the key role played by the *production force* of communication, as he terms what we identify as the moral resources related to the subjects' ability to understand one another and act jointly through their ability to reach agreements (García-Marzá 2008).

Ignoring this need for justification results in the exclusion of civil society institutions from any critical perspective, as well as any possibility of transformation and change that does not come from other spheres, via a legal route in the case of the state, and influence in the case of public opinion. As a result, the radicalism of this Habermasian proposal diminishes, as it does not encompass all the areas where power exists. To account for the presence and power of these actors for social change and transformation, we need a broader concept of civil society. With regard to Adela Cortina's approach, the participatory ideal opens up an approach to finding a solution, as it recognizes and emphasizes how "civil society is also a public dimension of society, where people can and should exercise their autonomy and defend generalizable interests through their exercise of significant social participation" (1993, 144). Once again, participation is a hallmark of the attribution of public responsibility to civil society. In the tradition of Tocqueville and Walzer, civil society is understood broadly as the network of social structures that are not directly dependent upon the state. The term brings together all the initiatives, capacities and resources used to make institutional arrangements, so long as they are the result of the unhindered, free and voluntary nature of individuals (Walzer 2008). The voluntary approach to establishing social arrangements defines moral validity and establishes its possible legitimacy or justice. Therefore, individuals' active participation—their capacity to make free, voluntary agreements—constitutes their condition of possibility. Here we encounter public opinion as a sphere where citizens can express themselves freely; civic and solidarity

associations and organizations; but also markets and organizations and institutions in the economic sphere.

From our perspective, within the spheres of civil society, we encounter particular interests (prestige, money), but we can also encounter common interests (professional, corporate, and so on) as well as general or universal interests (recognition, dignity, and so on). In our opinion, as is also encompassed by this proposed concept of civil society, the moral sphere is identified with generalizable interests but it is not closed in on itself as is the case for Habermas's formulation. In all these spheres of civil society, strategic and communicative actions are always combined with one another. Across all its actions and institutions, this integration is necessary in order to achieve the social good pursued (Apel 1982).

Of course, in Rawlsian terms we encounter an unequal distribution of basic resources in economic institutions, such as power and money, along with an asymmetrical distribution of burdens and benefits of social cooperation. However, we find the same in all areas of civil society and the state. The legitimacy of institutions consists of their free and voluntary acceptance of these asymmetrical conditions as the best way to achieve the specific social good that defines them.

The definition of civil society thus becomes a normative criterion; a *horizon of meaning*; whose validity—its quality—is measured by the distance that separates us from the inclusion and agreement of all those affected. A validity criterion is used to measure this distance; it is in the conditions that enable this acceptance, and which require equal conditions of participation, that such asymmetries *cannot* exist. Once more, the characteristic of civil society is the *primacy of* communicative logic, but its *exclusivity* is renounced. Communicative action, understanding and possible agreement, must be integrated with the specific good that characterizes each area or sphere of civil society. Their legitimacy and, empirically, the trust it deserves depends on the distance between the integration achieved and its possible free and equal acceptance by all those involved and affected (García-Marzá 2003).

The broad definition of civil society proposed here continues to rely on discourse ethics and the procedural conditions for reaching agreements. The only difference is that the reconstructive analysis of the participants' intuitive knowledge is not confined to the political sphere. We are also participants, who are involved in and affected by the formation and maintenance of institutional arrangements that are not strictly political, and nothing prevents us from applying the same reconstructive methodology to members of civil society, such as workers in a company, teachers or students at a university, members of a trade

union, patients in a hospital, and so on. Briefly, it is not the exclusivity of communicative action that legitimizes civil society, but its *primacy* when understanding and applying any form of coordination of action.

To make this *moral strategy*, as Apel would call it, explicit, we need a methodology that allows us to reconstruct the ethical foundations of trust in each of the political and non-political spheres that constitute the democratic system. Cortina proposes her own methodology for approaching the moral knowledge inherent in the plurality of practices that characterizes civil society: "The structure of applied ethics is neither deductive nor inductive, but benefits from the circularity of a critical hermeneutic, as it is within the different areas of social life that we detect an ethical principle as a backdrop, a principle that is modulated in very different ways depending on the area in which we find ourselves" (Cortina 1993, 174).

The *critical perspective* is established through reflective argumentation, in our case through the analysis and reconstruction of the conditions of possibility that allow us to speak of autonomy and equality: free and voluntary agreements. If these conditions do not form part of the definition of civil society itself, we have no critical perspective and are therefore unable to distinguish between a good and a bad civil society; between the justice and injustice of its institutions.

The *hermeneutic perspective* emerges when we want to engage with the diversity and plurality of the different spheres of civil society. Here the moral demand for equal and effective participation must be combined with the specific logic pertaining to each of these spheres, as well as the strategies needed to achieve the specific good which it is expected will be provided. The fact that strategic action is necessary in these different areas does not mean that communicative action, dialogue and subsequent agreement have no role to play in companies, universities, foundations, trade unions, etc. In all these spheres, both logics are entwined in a balance defined by the social good being sought, on the one hand, and the material conditions and available resources, on the other. The justice of this balance, always and ultimately depends on the possible agreement of all parties involved and affected in a given institutional power structure.

Through the application of this critical hermeneutic, we can ensure civil society institutions are held accountable for their power and for their capacity to include all those involved in defining and satisfying the interests at stake, while also achieving the real and effective participation of the different interest groups making up each type of institution. The possibility of agreement continues to be the normative horizon that grants these institutions' legitimacy, while free and equal participation are the only way to achieve it. But how is this participation

possible?

A systemic approach to deliberative democracy shows that the practical realization of this demand for participation takes place both at the macro level—public opinion and social movements, and so on—and at the micro level—micro-publics, deliberative forums, etc. However, we do not find approaches that we could call *meso-deliberative* and that are responsible for developing the institutionalist perspective within this deliberative model (García-Marzá 2016). It is precisely this intermediate level that we are interested in highlighting if we are to continue with the idea of a two-way democracy, as we can thereby manage to reflect on the participation possible within civil society institutions, for example, within large technological corporations (Hendricks 2006). It is not a question of replacing representation, but of seeking new spaces for dialogue and agreement: spaces of trust within civil society institutions such as companies, universities, churches, and so on, each with their different degree of institutionalization. Such an approach must integrate both the normative assumptions defining mutual respect and free and equal deliberation when searching for agreements and their potential for factic undertaking, as well as taking into account the available capacities, motivations, resources and constraints. Needless to say, this is a power: a capacity to satisfy and define interests that goes beyond the state and that must be investigated from a reconstruction of its meaning and validity in each of the spheres of civil society.

11.3 A Participatory Model of Deliberative Democracy

Understanding deliberative democracy as a *two-way democracy* implies incorporating into its inherent reflection the power of resources that are specific to civil society and that cannot be reduced to mere influence. While the *state* comprises the institutions with the legal power to bind our decisions and behaviour, and the structure we call the political sphere (government and political representation; legislative, executive, judicial power; public administration; political parties, security forces; etc.), *civil society* comprises all institutions that do not depend directly on political power, and which follow their own logic in order to achieve the basic social good they provide (universities, companies, hospitals, chambers of commerce, media, social movements, civic associations, and so on). In both cases, the free, voluntary acceptance reached after an agreement attained under equal conditions of participation defines their moral validity (Keane 2013)

According to this critical hermeneutic, applied ethics can be understood as a structuring of civil society, but only if they include a necessary institutionalist perspective. In our view, from this perspective, applied ethics and democracy theory are integrated into an *ethics of institutional design,* which is charged with designing and redesigning the different institutions of civil society so that their structure enables and enhances spaces of participation for the use of this moral knowledge within its institutions. A renewed ethics of discourse must incorporate this broad vision of civil society if it does not wish to silence the spaces where injustice is currently most frequently produced and reproduced; that is, if it does not want to forego its critical function (Cortina et al. 2008).

It was a mistake of participationist theorists in the 1960s to identify the social and the political, as this reduced public responsibility to state responsibility. It also ignored the logic of each sphere and considered that their forms of functioning, and therefore of participation, could be assimilated to the political sphere. Today's participatory democracy does not seek to replicate the systems of election and electoral decision-making within the institutions and organizations that make up civil society. This is neither possible nor desirable, as has been shown, for example, by the failed experiences of industrial democracy or, even more recently, the relationship between municipal democracy and the corruption that resulted from it (Warren 1999). The demand for participation must always be mediated with the meaning and logic that defines the operation of each institution.

Nor are we saying that institutions' responsibility begins and ends with their contribution to political power. Civil society has a value in itself; it is not a "secondary circuit" for politics. An ethics of institutions is instead addressed to the organizations themselves as *institutional actors*, and what we are interested in is their contribution *qua* institutions, in short, how they help to make explicit the conditions of the agreement that legitimizes them as such. In the absence of this perspective, it is difficult, if not impossible, to assess their potential for social transformation.

We know perfectly well that a direct application of the ethics of discourse to democracy is neither possible nor desirable, as it would only lead to a false moralization of politics, which would be essential sustenance for the technopopulism discussed above. However, in our opinion, we would have a chance to update this model if, firstly, we stopped identifying participatory democracy with direct democracy and specified what *meaningful, equal and effective participation* means in current contexts; and, secondly, if we stopped confusing representative democracy with democratic representation and positioned participation (as opposed to representation; although this is still *a*

form of participation) as the normative core of democratic theory.

As we have endeavoured to demonstrate, a participatory democracy is a moral aspiration, a moral demand, but it can and must also be a way of organizing our collective life. To this end, a democratic theory must have an ethics of institutional design that enables the creation of spaces for participation, which are understood as *new public spaces* within institutions. These should be spaces for deliberating and deciding on interests that are capable of being generalized; of capturing and generating public opinion within organizations; of monitoring their public accountability; and, finally, of establishing their claim to justice. They are spaces where the different interests at stake are represented, using different mechanisms, such as trust or even the drawing of lots, for representation. An institutionalist perspective is needed to undertake this participatory updating of deliberative democracy, or, more specifically, an ethics of institutional design that allow us to operationalize the gap between the demand for participation and real, affective participation. In dialogic digital ethics this bridge is built with ethical principles and, in particular, the principle of explicability, which is understood in terms of the Kantian principle of publicity as the sum of transparency and participation (García-Marzá 2012).

The current interest in institutional design and, ultimately, in rethinking the meaning and legitimacy of the institutions that sustain civil society, is a response to a situation of dissatisfaction with the problems afflicting our democratic systems and the theories that should help us to understand and solve them (Goodin 1998). An ethics of institutions aims to combine the efforts of applied ethics and critical reflection on democracy in order to build a public argument capable of responding to the resources of civil society; of building a civil and not only a political public space. As a critical outlook the ethical perspective must support the idea of empowering subjects and their institutions to recover and assume their share of responsibility for the public sphere.

Civil society is not political society, but it is a decisive part of democracy. Not all social relationships are political relationships, but if there are power relationships within them, they must be part of a democracy understood as a form of society: a means of solving public problems. If companies, universities or hospitals have the power and capacity for this purpose, it is very difficult to understand why they are not part of the web of institutions that make up the democratic system. In my opinion, this distribution of responsibility for public affairs is the best way to curb the disastrous but growing algorithmic democracy, and likewise the loss of individual and collective autonomy that it entails (Calvo 2018).

It is clear that the representative system is losing credibility today and its

structures of participation need to be deepened, above all with regard to political parties and the separation of powers. However, an ethics of institutional design has as its basic priority the institutions of civil society simply because it is there that our moral knowledge is transformed into moral resources: into an ethical capital capable of empowering subjects in their different spheres of action, curbing the consensual loss of autonomy entailed by all datafication. It is precisely because within institutions that participation is focused on specific causes; on actions anchored in the contexts of the lifeworld directly related to our interests, our work, our health, our families, and so on. This is the best way to recover participation and, with it, to overcome contemporary demoralization: by linking participation with moral emotions by designing spaces for participation within institutions, and among those institutions committed to managing their responsibility as agents of justice. The inclusion of civil society in all its complexity and heterogeneity, along with its strong institutionalist tone, is a promising way forward for achieving "people-centred" technologies.

The concept of civil society, like any category that attempts to define a social reality, is far from uniform. From its origins, this term has always been defined in relation to the state, precisely as the network of social structures that are not directly dependent on it (Keane 1992). This does not mean they are two separate or separable spheres. Civil society cannot be conceived of without the state, as the latter's legal framework constitutes the basic support for the operation of its own mechanisms for interaction. Civil society is independent of the state in terms of its initiatives, its objectives and the means or resources it has at its disposal to achieve or attain them. As we have discussed, the problem arises when, within this new scenario, the extension of global institutions does not coincide with the legal framework. Working from these coordinates, civil society always refers to, as Walzer stated, both the space encompassed by non-coercive human associations and the network of relationships created for the defence of the family, religion, interests and ideologies that spread across this space (Walzer 1991). The central axis of any definition of civil society seemingly revolves around this non-coactive character of its practices and institutions. Since Tocqueville, it has been insistently stated that this sphere is linked to the active participation of individuals; to their capacity to make free and voluntary agreements to satisfy their interests and jointly resolve problems. Encompassed within this individual autonomy and the freedom to establish agreements lies the emancipatory force of civil society, along with the capacity to halt any violence, whether explicit or implicit (Kaldor 2020).

In order to include these spheres of action, the power that underlies them and the consequent legitimacy they require, we must bear in mind that these

demands for participation, as well as the agreement that communicative action entails, could lead to a paralysis of the logic of the economy itself: that is, the requirements of the market's efficiency and competitiveness (Cohen and Arato 1994). These requirements that seem to demand both an asymmetry in the conditions of participation and an unequal distribution of the results of the coordination of action. Secondly, within these areas, we are moving among particular, rather than general interests, where at most we can speak of strategies and negotiations as possibilities for conflict resolution.

Yet, if these motives are ultimately the unequal conditions for participation, the need for strategic action and the medium of money, the question is obvious: which institutions do not require a degree of integration of all these elements with communicative action? Do not churches, social movements, civic and solidarity associations require the medium of money? Are not the media, many of which are now in the hands of investment funds, businesses? Or vice versa, can the economy and its institutions be explained without talking about values, precepts, reciprocity, trust, and so on? Aside from the exclusion of economic institutions, the theory of democracy thus loses one of its basic pillars, as it excludes from analysis the spheres of action in which the vast majority of social relationships are established, and in which economic and social power is produced and reproduced (Calvo 2018).

11.4 Ethical Principles for Institutional Design

The concept of civil society, like the concept of democracy, can be nothing other than normative, or include a normative dimension if we want it to have one, because we use it to construct the same social reality, and it should be remembered that this reality is not independent of the subjects making it up. In our case, it is a normative concept of civil society that emphasizes the idea of participation and agreement, whatever the extent of any problems or the institutions created to solve them. It is, therefore, a concept of a global civil society that can undertake the task of overcoming the state framework because it has its own mechanisms for coordinating action, referred to here as moral resources, which are derived from the capacity to establish communicative actions and the moral commitments they entail. The boundaries of civil society —its frontiers—are determined solely and exclusively by the dimension of the conflicts to be resolved in a dialogic and reasoned manner (Keane 2013). These conflicts are not limited to generalizable interests, but also involve group or common interests, as well as private interests. As a normative concept, civil society insists that everyone has an equal chance to participate in this search for

contracts if they so wish. To put it in Habermasian terminology: communicative action is a priority. This is none other than the moral horizon for action and the criterion that legitimizes the achievements of civil society.

The critical perspective is determined by this normative character, supported by the horizon of equal participation of all those affected and involved in the search for agreements. The idea of agreement and the distance that separates us thus become a criterion of validity for assessing the legitimacy of an institution and, factically, a rational basis for building trust. The idea of agreement is not some fictitious ideal invented by theory, it derives from the reconstruction of our own intuitive knowledge. It therefore has a factic reality, i.e., it provides us with cognitive content, which allows us to assess the legitimacy of the institution and, consequently, its affiliation to civil society. It is not the achievement of the social good sought, be it efficiency, health, defence, education, or whatever that determines membership to civil society, but rather the possible consensus of all parties involved concerning whether that goal has been achieved. Hence, the major issue when it comes to establish this proceduralist criterion of justice is initially to reconstruct the *guiding principles*, the conditions of possibility, for the design of these spaces.

These principles are not merely inventive, nor are they the outcome of a comparison of what "really exists" in our current context, as we have seen groups of experts do. It is rather the application of what we have called critical hermeneutics to meaning, validity and moral credibility of institutions. According to Offe (2019), this credibility depends on two variables, two types of reasons to justify its power:

– In the *functional dimension*, institutions are seen as mere instruments that individuals construct to solve basic problems and achieve certain social aims and goals. They are mechanisms for the resolution of conflicts of action and their arguments therefore relate to the efficiency and proper operation of the institutions. They depend on the achievement of the social good that justifies their creation and maintenance and they have to fulfil the task or mission entrusted to them, which means they are compatible with the supply of resources they depend on. This instrumental understanding concerns the implementation of objectives and the control of resources, as well as making things happen: its logic is of a strategic-instrumental order. The rules relate to this division of labour and are based on a dense web of incentives and punishments. Essentially, it a strategic dimension that contains the reasons concerning the achievement of the good and how it has been achieved, along with the optimization of resources.

– In the *normative dimension*, institutions are seen as producers of meaning, and here their arguments concern communication and agreement, as well as the reciprocal acceptance of the claim to validity raised by the values, precepts, rights and duties established by the institution. Reference is made to the hegemonic or prescriptive function of the institutions in charge of the justification discourses. These establish identity and a common framework for action, as well as what is essentially the meaning of the institution. In this way, they manage to define what is normal, what can be expected, what can be trusted, which rights and duties correspond to which position and what is meaningful in the social sphere or community. They thereby render an institution valid. They address the agreement or understanding between the parties, whether or not this is assumed, and to reach such an agreement communicative logic does not suffice. An institution survives if it passes the double test of being meaningful and appropriate. Institutions can thus be classified somewhere between social norms and norm-driven action, and strategic or rational action towards a specific aim (Offe 2019).

However, most institutionalist theories fail to address this dual character. In many of these theories the difference between "good" and "bad" design continues to be reduced to either a conventional harmony with or adaptation to an environment, or else to a strategic calculation of the achievements made by a particular interest group (Goodin 1998). In all these cases, the need for justification goes no further than social acceptance or efficiency, while ignoring the fact that behind every decision-making strategy; every distribution of burdens and benefits; there is a shared social meaning that makes them possible, giving them meaning and reality. This meaning is not unrelated to the claims of validity or legitimacy made by every institution.

Reality demonstrates the failures of these unilateral approaches. For example, the "intentional" adoption of institutions for the purpose of promoting some economic objective can ruin their intended effects if the institutions are transplanted as a formal set of rules and procedures rather than cultural values and norms, which provide the reasons that generate consent to those rules. As has happened on many occasions at both market and company level, these "institutional grafts" frequently fail, as we will discuss with regard to Google's Ethics Committee. On the other hand, there is no point in clinging on to the current institutional structure whose self-understanding we have grown up with when its effectiveness at solving basic problems has been greatly diminished, for example, due to the development of digital technologies such as e-democracy or cyberdemocracy.

Such examples highlight the need to consider both logics for addressing institutional change and transformation: the fact that institutions do not constitute a closed and definitive social phenomenon; instead, we are always confronted by a *process* within which institutional design acquires its meaning and *raison d'être*. This dynamic character of the institutions is only the expression of the difficult balance between *efficiency and justice* inevitably encountered in every institution. From the participant's perspective, institutions do not form an unchanging social *faktum*, but rather a process through which social practices and the institutions in which they are structured acquire meaning and stability.

By applying this dynamic view of institutions one can easily perceive the partiality of any approach centred on a neutral observer who reports on the effectiveness or level of success achieved by the institution through the contribution of its social benefits. If we turn from the observer's to the participant's perspective, it is no longer possible to separate the question of what we have achieved from the question of how we have achieved it. It is not possible to give an account of the stability attained or the security and coordination mechanisms, or therefore of the selection and transformation of preferences (desires and needs), unless we refer to the justification or legitimization of these systems of rules: the understanding or agreement which they are based upon. Institutional design is not only a pragmatic question of what it means to achieve certain ends, nor a merely psychological question of how to motivate individuals. It is also a moral question that concerns the fairness of the relationship that is established.

In short, the justification or legitimacy of an institution does not depend only on whether it has achieved the proposed social aim or benefit, yet this question cannot be addressed separately from the issue of the consent or agreement of all those affected. We should not seek an institution's morality through the calculation of its effects, nor whether it has achieved the proposed end, but instead with regard to an agreement that reflects the reciprocal recognition of all those involved and/or affected. The first question is monological, regardless of which interest group carries it out; the second is dialogic, as it admits only an outcome that is the result of deliberation. *Institutional morality thus refers to the conditions underpinning the freedom and voluntary nature assumed by civil society institutions; in other words, that which underpins the deliberation and subsequent agreement justifying these institutions* (García-Marzá 2019).

The ethical perspective on institutional design is concerned with the conditions required for the production and reproduction of this moral agreement, which underpins the belief in an institution's legitimacy. Its aim is to

demonstrate that moral reasons provide the foundation for this rational conviction concerning the validity or justice of institutions, and that this rational motivation is not explicable only by the force of custom, rites and traditions or in terms of expectations of their utility. It is these moral reasons that underlie any possible agreement, rather than its contents, even when we start from the *de facto* asymmetrical conditions of burdens and benefits, as is the case for most civil society institutions. Here, too, we are talking about more or less fair or morally correct commitments and negotiations. It is this moral knowledge that we continually use as members of an institution that concerns us.

The moral quality of an institution, along with it its capacity to generate trust and create spaces for action where it is possible to act morally, would thus be defined by its willingness to reach an agreement or consent with all parties involved. However, in order to talk about institutional design, we now need to convert this moral dimension into a set of principles that will be operational in this respect. The real challenge lies in presenting this operationality without relinquishing any of its moral obligation: its unconditionality. It should be recalled that it is this unconditional character that enables moral knowledge to perform its function.

On the basis of this reconstruction, an institution's legitimacy is not reduced to its functionality, nor to its morality. Instead, it is a network of discourses or reasons, whose decisive factor is the free play between its component parts. The key is to understand this legitimacy as a process which these different discourses become implicated in, and where morality undertakes to define free access and equal participation based on free acceptance. The agreement upon which the legitimacy of institutions depends upon refers to this deliberation, according to which moral reasons are combined with compromises and strategies. However, this is why we say that effectiveness is not at odds with morality, nor are they two separate issues. They are two moments in a process through which institutions achieve justification. We speak of reflexive arguments because this moral dimension becomes concrete, precisely, as Dryzek would say, as part of the meaning and possibility of the institution itself (Dryzek 2001).

In order for these discourses to function and for reasons to become reasons to act, to generate what Habermas calls *communicative power*, it is necessary for the structure of these deliberations to allow free access and movement between these discourses; in other words, the freedom to participate on equal terms in all of them. It is only from these perfect conditions of symmetry and equality, and not based on strategy, force or violence, that there can be a rationally motivated consensus. In other words, not merely any deliberation would be enough to justify the agreement, but the agreement can be considered fair when it is made

under conditions that ensure freedom and an equality of participation and thus the agreement's voluntary nature. It is on the basis of the idea of the conditions underlying a free and equal agreement that we can derive the principles we can be reconstructed for institutional design.

We are facing the same normative (ideal) conditions that underlie the meaning of the political public sphere, but these now *extend to all spaces where power exists: to the civil spheres, including the corporate public sphere, which are essential for the credibility and legitimacy of technological platforms and the other institutions of civil society.*

Drawing on various works by Habermas, we can summarize these conditions as two basic principles:

– *Principle of inclusivity:* institutional regulation must grant the possibility to all those, actually or potentially involved and affected to participate in all decisions that affect them, as well as including all their interests in the deliberation process.
– *Principle of reciprocity:* all participants should have equal rights of participation, so as to ensure a full equality of opportunity for deliberation and subsequent decision-making. All interests should in principle be open to discussion.

However, these principles only tell us what a free and equal agreement means; they tell us nothing about the necessary mediation of this free and equal participation with the effective achievement of the social good pursued by the institution. From this point of view, it is obvious that these conditions are far from being met in most of the institutional arrangements that make up both the state and civil society. In the case of civil society, an ethics of institutions must not only be aware of this fact but take this non-compliance as its point of departure. The question that arises is quite clear: how should these conditions be applied to civil society; to social arrangements expressly designed so that the hierarchy and the unequal distribution of burdens and benefits achieve the desired social good?

In order to undertake a reflection on the moral dimension of institutional design, we must first analyse how these conditions are institutionalized in contexts that, by definition, seemingly contradict them; in other words, they deviate from the moral horizon established by the first two principles. The claim to legitimacy—the need to justify its power structure—remains the same. The question now is how to account for it. Two options are open to us in this regard.

The first of these, to which Habermas has contributed, consists of leaving all the institutions of civil society not concerned with generalizable interests outside

the democratic system, and instead placing them in the catch-all category of systemic analysis. The moral dimension's contribution to the design and transformation of these institutions is lost, beyond the necessary but insufficient contribution of public opinion. It is as if we thought the market required no more rules and no more legitimization than that deriving from the law of supply and demand; or that in order to obtain good health professionals, a health law would suffice.

The second option is the one we advocate. This consists of acknowledging that these two principles must be integrated into a third principle that corresponds to the pragmatic dimension of the institutions:

– *Principle of effectiveness:* institutional structuring must pursue the achievement of the specific good that characterizes each social practice and its institutions as basic pillars. However, efficiency and robustness are values that cannot be separated from what we expect from each institution and each form of technology.

The idea is that, while it is not possible to put forward the premises under which it might be rational in specific contexts and institutions for people to consent to an agreement under unequal conditions, it is possible to establish the necessary conditions for that institution to publicize the institutional structure that has been created. To this end, institutions must "show" (make public) the efforts achieved by following the specific path they have taken towards the intended agreement. Therefore, this condition of publicity refers not only to the results achieved by the institution, but also, and basically, to the institutional rules and norms that enable participation and agreement concerning the achievement of that good.

Institutions can clearly only fulfil their mission of aligning behaviour with social relations to the extent that they are known, for this is the only way that they can shape the expectations and will of the parties involved. This knowledge is also public because it affects the reciprocal knowledge of others: it is not possible to form this agreement without each member knowing what the rules demand of them, as well as the consequences of following these rules, and, especially, without the rest of those involved or affected knowing that this is known. However, there is still one more step to be taken, as the agreement reached must also be made public. Without this publicity, there is no possibility of stability and continuity: the true pillars of the institutions. Hence, insistence must be placed on the public nature of its regulatory structure, and the agreement to which this publicity refers. The need for public knowledge and the intended agreement are crucially important to understanding the importance of the

Kantian principle of publicity.

Ultimately, this public nature of the agreement is not only necessary for its effectiveness, but it also ensures it is just. The closer the institutional structure comes to the idea of agreement, the less the institution will need to resort to secrecy, manipulation or outright violence. It is injustice that leads to the use of force and not the other way around as we tend to think. The principle could be interpreted as follows:

– *Principle of explicability*: the institution must be designed in such a way that we are aware of the level achieved through meeting the conditions that lead to or can lead to an agreement that legitimizes them.

We have already seen how the principle of explicability refers precisely to making public the design, as well as the development and calculation, of the consequences deriving from all forms of use of Artificial intelligence across the different processes that make up democracy (García-Marzá 2012). Technical competence is no impediment to the participation that these design principles call for, and if it is, this is due to instrumentalization or bad design. If we cannot explain what lies within an algorithm, the conditions for its development or its consequences, its technological application must be stopped if we are to build trust in digital practice. Our proposal is that the design of institutions should follow these principles, which are based on the conditions of possibility for all agreements and compromises.

Drawing on this proposal for expanding the strict limits of deliberative democracy—in other words, through the protagonism of civil society—the consideration of civil society as a space where social agreements are generated for the possible resolution of conflicts of action is maintained, thereby halting the progress of an algorithmic culture that is gaining more and more ground and thus defining what is possible and desirable. This amounts to the recovery of face-to-face, real, affective participation, in order to recover our autonomy within a digital citizenship. To achieve this requires designing spaces for participation within the institutions that make up civil society, spaces for bringing together the moral capacities of individuals in order to assume a responsible commitment to cooperation and collaboration. The incompatibility between aggregative or deliberative democracy and algorithmic democracy derives from a reality that has been both demonstrated and concealed a thousand times: no institution, no matter how powerful and no matter how many digital technologies it possesses and dominates, can be maintained without social credibility, and this moral capital is held exclusively by civil society (García-Marzá 2012).

As a result, it is necessary to think, debate and jointly design trustworthy spaces that develop institutions' capacity to empower subjects, and that are institutionally ordered, flexible and open spaces, where subjects can use their capacities for understanding and agreement. Any possible response requires an institutionalist perspective, to provide principles for the design and redesign of institutions that help us to think of new and better political, social and economic institutions, assumed based on our autonomy and our capacity to jointly decide the criteria for justice that are closer to the idea of justice. Digital technologies can help us to create and maintain these spaces of presence, but not to replace them with virtual realities, whose ultimate aim is to control and commercialize our opinion and our will, and thus hide the injustice that really exists.

From an ethical perspective, no "recipes" for institutional design can or should be given. The role of ethics is much more modest and has to do with the need for self-justification inherent to any claim to legitimacy, along with the principle of explicability. Institutional morality refers to something that comes before the calculation of consequences, situated amid the conditions that make deliberation and agreement on this calculation possible. If we manage to channel participation through the institutions that make up the socio-technical context of algorithmic democracy, we will have a complementary path to legality; a path that does not rely on coercion but on self-regulation and the public. We will also have an explicit commitment to give an account of how and why its power has been used and, therefore, its capacity for change and social transformation. In short, an explanation of its claim to be just.

Bibliography

Achen, C.H., and L.M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton: Princeton University Press.
[Crossref]

Apel, K.O. 1982. Normative Ethics and Strategic Rationality: The Philosophical Problem of a Political Ethics. *Graduate Faculty Philosophy Journal* 9 (1): 81–107.
[Crossref]

Calvo, P. 2018. *The Cordial Economy-Ethics, Recognition and Reciprocity*. Cham: Springer.
[Crossref]

Chambers, S. 2002. A Critical Theory of Civil Society. In *Alternative Conceptions of Civil Society*, ed. S. Chambers and W. Kymlicka. Princeton: Princeton University Press.
[Crossref]

Cohen, J.L., and A. Arato. 1994. *Civil Society and Political Theory*. Cambridge: MIT Press.

Cortina, A., ed. 1993. *Ética aplicada y democracia radical*. Madrid: Tecnos.

Cortina, A.C., J. Conill, and D. García-Marzá, eds. 2008. *Public Reason and Applied Ethics: The Ways of Practical Reason in a Pluralist Society*. Ashgate Publishing Ltd.

Dryzek, J. 2001. Legitimacy and Economy in Deliberative Democracy. *Political Theory* 29 (5, October): 651–669.
[Crossref]

Garcia-Marzá, D. 1992. *Ética de la justícia. Habermas y la ética discursiva*. Madrid: Tecnos.

———. 2003. *Ética empresarial: del diálogo a la confianza*. Madrid: Trotta.

———. 2008. Sociedad civil: una concepción radical. *Recerca: revista de pensament i anàlisi*: 27–46.

———. 2012. Kant's Principle of Publicity. *Kant-Studien. Philosophische Zeitschrift der Kant-Gesellsschaft* 103: 96–113.

———. 2016. Enfoques mesodeliberativos: sobre la articulación institucional en las democracias deliberativas. *Isegoría* 54: 147–170.
[Crossref]

———. 2019. Repensar la democracia. Estrategia moral y perspectiva crítica en K.O. Apel. *Daimon* 78: 75–89.
[Crossref]

Goodin, R.E., ed. 1998. *The Theory of Institutional Design*. Cambridge: Cambridge University Press.

———. 2008. *Innovating Democracy: Democratic Theory and Practice After the Deliberative Turn*. Oxford: Oxford University Press.
[Crossref]

Habemas, J. 1992. *Faktizität und Geltung*. Suhrkamp: Frankfurt.

Hendriks, C.M. 2006. Integrated Deliberation: Reconciling Civil Society's Dual Role in Deliberative Democracy. *Political Studies* 54 (3): 486–508.
[Crossref]

Honneth, A. 1993. *The Critique of Power: Reflective Stages in a Critical Social Theory*. Cambridge: MIT Press.

Kaldor, M. 2020. Global Civil Society: An Answer to War. In *The New Social Theory Reader*, 252–259. London: Routledge.
[Crossref]

Keane, John. 1992. *Democracia y sociedad civil*. Madrid: Alianza Universidad.

———. 2013. *Civil Society: Old Images, New Visions*. New York: Wiley.

Maschewski, Felix, and Nosthoff, Anna-Verena. 2021. Der plattformökonomische Infrastrukturwandel der Öffentlichkeit: Facebook und Cambridge Analytica revisited. In *Ein neuer Strukturwandel der Öffentlichkeit?*, Hrsg. Martin Seeliger, and Sebastian Sevignani, 320–341. Frankfurt, Suhrkamp Verlag.

https://doi.org/10.5771/9783748912187-320.

Offe, C. 2019. Political Institutions and Social Power: Conceptual Explorations (2006). In *Institutionen, Normen, Bürgertugenden*, 3–22. Wiesbaden: Springer VS.
[Crossref]

Vitale, D. 2006. Between Deliberative and Participatory Democracy: A Contribution on Habermas. *Philosophy & Social Criticism 32* (6): 739–766.
[Crossref]

Walzer, M. 1991. The Idea of Civil Society: A Path to Social Reconstruction. *Dissent* 39 (spring): 293–304.

———. 2008. *Spheres of Justice: A Defense of Pluralism and Equality*. Basic Books.

Warren, M.E. 1999. *Democracy and Trust*. Cambridge: Cambridge University Press.
[Crossref]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
D. García-Marzá, P. Calvo, *Algorithmic Democracy*, Philosophy and Politics - Critical Explorations 29
https://doi.org/10.1007/978-3-031-53015-9_12

12. Institutional Design for the Development of Algorithmic Democracy

Domingo García-Marzá[1] ✉ and Patrici Calvo[2]
(1) Universitat Jaume I of Castellón, Castelló de la Plana, Spain
(2) Universitat Jaume I of Castellón, Castellón de la Plana, Spain

Civil society makes room for all these aims and includes all the resulting associations, by virtue of their free and consensual character. This means that it reaches politics as well as the multitude of social activities …This is an extended definition; there are theorists who would exclude whatever goes on in the marketplace (and sometimes also in the political arena) from their account of civil society, perhaps because they do not think of such goings on as "civil". But I see no good reason for the exclusion. If unions are to be included, as they standardly are, then why not the companies with which they are regularly engaged in negotiations over money and job security? To be sure, giant corporations sometimes exercise statelike, coercive, or near-coercive powers, but that seems a better reason for reforming and restraining the corporation than for excluding it from our definition of civil society.
Michael Walzer, Equality and civil society (2002)

Abstract
This last chapter presents several examples of how it is possible to create public spheres within the institutions of civil society, especially in large technological or big tech platforms. It is not possible to leave civil society and its institutions out of the democratic system, precisely because it is the space where both meaning and the power to construct it are produced and reproduced. We have seen that today, this power is in the hands of algorithmic colonization. In short, we will give an example of how a bidirectional democracy can stop this colonization from civil society. To achieve this, we will begin by considering civil society institutions as agents of justice and propose a new distribution of

responsibility, one capable of confronting the power wielded by the technological giants (Big Tech). Secondly, we address the validity and efficacy of what are known as voluntary self-regulation mechanisms. Thirdly, we will analyse two possible scenarios, within what we will call an ethical infrastructure that guarantees the conditions for equal participation of all those concerned. Finally, we will discuss the construction, legitimacy and efficiency of these new civil public spheres.

Keywords Algorithmic colonization – Civil society institutions – Deliberative democracy – Dialogical ethics – Big tech – Spheres for participation – Spaces for trust

12.1 Civil Society Institutions as Agents of Justice

Trust is a well-studied concept, and it is precisely in trust, and how it is generated and maintained that we can best appreciate the role of an ethic of democracy within the design of all types of institutions, both state and civil society. We trust a person or institution because we have information about the possible options and the consequences of any decision; we know our interlocutor's skills and competences, as well as the obligations arising from the action to be taken. Trusting someone always implies an *expectation of* reciprocity and with it a risk, since it always presupposes our interlocutor's goodwill; this is why we are vulnerable to the limits of this goodwill and need to guarantee our action's continuity. In our proposal, this involves creating spaces for participation where we have reason to expect a certain form of behaviour; spaces where understanding and agreement are possible, where moral commitment can be developed and where strategy is subsumed within an ethical framework (García-Marzá 2011).

Based on this trust-building perspective, it may be asked why the function and efficacy of moral knowledge does not imply adopting a utilitarian point of view. For an ethics of discourse, such as the one upheld here, the criterion that makes an action, practice or institution morally correct is not its social utility. It is rather the opposite: the consequences and effects achieved do not define the moral value of the action, but rather reciprocal recognition, the respect for the dignity we all deserve. It is this moral value that becomes a motivation for action, while the expectation that it will happen, becomes a reason for trust. Let us remember that an action is not morally right because it enables us to achieve this or that end, however important or decisive these may be. It is right or just because we have acted with respect for the rights of others, because we are

convinced that others are ends and not means for us.

The great challenge for an ethics of democracy consists precisely of reconstructing and using this moral capital across the different spheres of civil society (Offe 2004). Its basic task is to transform this moral knowledge both into the necessary virtues that underlie all kinds of social practices and the ethical design of our institutions. To achieve this, in earlier chapters we have discussed theories of institutional design and distinguished between individual and institutional responsibility (Goodin 1998). They are complementary and refer to one another, but they are analytically distinguishable. Applied ethics tends to dismiss this level, and overlooks the fact that vulnerability not only affects people, but also institutions. Building trust is necessary for reputation and reputation is necessary for credibility, all the more so in global contexts such as today's, where legal regulation can do little to resolve complex, accelerated and global conflicts.

When we talk about trusting a technology, what we really mean is trusting that the company which has developed it has implemented reliable processes for security, oversight, transparency, and so on (Green 2021a). A company, like any other institution, whether part of the state or civil society, requires social legitimacy and it needs to inspire trust to be able to operate. The reasons given in support of their social credibility always refer to the contribution made to society, the social response provided, and the impact generated. However, moral validity derives from the possible agreement of all those affected and involved by this impact, and not the utilitarian calculation reached by experts or their algorithms. The contribution made to the development of the communities with which they engage is an important part of this social response and one of the basic aspects that renders their activity credible and justifiable. This is all the more relevant in the case of the major technological corporations, who are reducing culture, politics and the economy to the buying and selling of data in the name of the highest bidder, whether that be the state or the market. To say that technology companies and research centres are not accountable for what they do is the shortest route to abandoning the democratic system and increasing algorithmic datafication, and thereby the oblivion of justice concealed beneath the cloak of neutrality. The digital revolution is no longer the utopia we had hoped for; it is the greatest danger we face in this first quarter of the century. The question is simple: are companies and, in our case, large technology corporations, part of the system of institutions that make up democracy, or does their power, and therefore their responsibility, lie beyond it, beyond any appeal to justice?

Ethics of democracy understand the contribution of business, the creation of

shared value, from a critical perspective, in order to account for the role business does, can and should play as an *agent of justice*. At present they are actors who wield unprecedented power and are capable of influencing the meaning and direction of the entire democratic system, yet they are reluctant to acknowledge their share of responsibility.

Despite this asymmetry between *power and responsibility*, which is even more profound in the case of the technological platforms, studies devoted to the theory of democracy rarely engage in a systematic analysis of companies as agents of justice, change and social transformation and, in particular, they neglect how *institutional design* could allow us to balance the power and responsibility of their actions. This situation corresponds to a widespread understanding of companies as machines, unchangeable institutional arrangements, whose structure and functioning are accepted or rejected *en bloc*, independently of the will of those who form part of them; a view of business shared by advocates and critics of global business performance. Such views give the impression that the major technology companies are a natural phenomenon and could not exist in any other form.

The *institutional reification* of corporations facilitates their low visibility within democratic theories and, thereby the lack of a clear definition of their responsibility for the impact they generate, which ranges from the surveillance, control and commercialization of our data to the collapse of public opinion, and the restriction of freedom and increasing inequality. As result companies are seemingly *amoral* and are not held to account for the capacities they possess in order to solve, should they wish to, the problems they pose, so they can align themselves with people-centred technological development, for example. It does not require a great deal of intuition to perceive how the extent and depth of their impact is changing the relationship between citizens and the state, and shifting towards authoritarianism. Nevertheless, it is not accepted that they are part of the democratic system (Robbins and Henchke 2017).

This neutrality, Big Tech's apparent goodwill has not been able to conceal its excessive power and lack of responsibility. Today governments are monitoring them; while they provoke public mistrust. The opacity of their base algorithms; their disdain for privacy and intimacy; security problems; the excess of accumulated power; censorship policies; the construction of political will; precariousness employment, and so on, all obscure the benefits they could bring were they to believe even minimally in their commitment to serving the people (Martin 2020).

The question that underlies this line of thought is whether we can conceive of an institutional design that would enable these corporations to become "agents

of justice", accountable to the autonomy and free will that underpin the democratic system? Or are they only and inevitably *agents of injustice*? Essentially, all forms of democratic thought must present the major technology companies—the owners of the data—as global actors who must be held accountable for their actions. This will provide protection against the moral impunity with which they act (Bietti 2019).

12.2 Ethics and Regulation: Handing Public Engagement with Big Tech

Today, business provides a clear example of the new role played by civil society, it also illustrates the new distribution of responsibility for public affairs, and this is demonstrated by the direct call for its current involvement in the fulfilment of the Sustainable Development Goals. Companies are increasing their capacity for action and, along with it, their social responsibility, which is understood as the need to be accountable to society for their actions. However, the major technology companies seek to evade this responsibility by hijacking the word ethics and instrumentalizing the concept to the point of absurdity and ridicule. Indeed, they have done the most to generate mistrust in any application of ethics as a public commitment and management tool (Garcia-Marzá 2005; Bietti 2019). In fact, the concept of social responsibility dovetails with these new expectations, but instead they seem to be losing their capacity to generate trust. Today this concept is, so to speak, "under suspicion". Social responsibility has proven itself to be unequal to the problems and demands placed on business, or because it is simply a guise to facilitate and mask bad practices. *There is even talk of ethics washing*. Some critics have called for it to be abandoned, instead placing their trust in greater legal regulation (Calo 2017; Hagendorff 2020).

However, ethics refers to the moral knowledge we possess as people, which is the fruit of our condition and our history. It is not easily instrumentalized. Despite their multimillion-dollar efforts and their social media monopoly, they have managed neither to overcome the mistrust, nor dispel the discredit currently identified with them. There is no need to underline the enormous power they possess, it is enough to recall that they are capable of modifying state laws in their favour, evading taxes on their high profits, fabricating public opinion, commercializing our private life, and, among other issues, redefining the concept of precariousness, and so on. Their negative impact is compounded by their opaque management and the algorithms they work with, their lack of transparency and evident willingness to lie in the face of public opinion, which,

despite their efforts, continues to maintain a critical and suspicious stance towards them. At the same time, state authorities are moving forward with regulation, led by the European Union is leading the way. As is well known, legal coercion becomes most necessary when reciprocity and trust fail.

The result of this situation is *the instrumentalization of ethics* and its values for totally spurious and fraudulent purposes, the use and abuse of moral concepts, voluntary public commitment campaigns, self-regulation and governance mechanisms, all for the sole purpose of shareholder profit. It is hard to deny that the ethical language which Big Tech companies allude to merely provides an effective mechanism to curb governments' regulatory anxieties. Ethics are increasingly identified with corporate self-regulatory efforts and superficial demonstrations of ethical behaviour (Bietti 2020; Metzinger 2019). Criticism of the instrumental use of ethics by these centres of power has focused on three different fronts:

1. Ethics are seen as an acceptable facade for justifying deregulation, self-regulation and market-driven governance. Hence the self-interested adoption of appearances of ethical behaviour to counter a tarnished reputation.
2. Ethics are seen as a communication strategy—a veil hiding injustice and prevailing malpractices. They blur the impact of what is really going on under a cover of claims of greater effectiveness, and a contribution to the improvement and progress of humankind.
3. Ethics are limited to the individual sphere, at most to the professional sphere, as if it were the sole preserve of each individual's conscience, and were therefore at the mercy of arbitrariness and subjective will as to action to the growing power and influence of Big Tech companies". The tech giants we know today, once hailed as society's heroes, had been prepared for the coming techlash for several years now, but in 2018 the storm truly arrived. The term techlash, a hybrid of "technology" and "backlash", seems to have originated from the title of an article published by The Economist in November 2013. Data privacy, or rather, the lack of it, has played a central role in this technological clash; the Cambridge Analytica scandal, which saw data concerning millions of people harvested from Facebook and used for funded political campaigns to influence voters on both sides of the Atlantic, fundamentally undermined public confidence in the ethics of the tech industry and its ability to manage what it has created. Neither Facebook CEO Mark Zuckerberg's tearful testimony before the US Senate, nor

> Europe's stringent General Data Protection Regulation (GDPR) laws have been able to assuage concerns about data privacy and its far-reaching implications for our democracy. In contrast, they have arguably served to increase public concern about all things data-related, while undermining any presumption of goodwill, along with the idea that these companies operate to serve the people (Green 2021b). The adoption of the word techlash by the Oxford dictionary signals how the issue is now at the forefront of public consciousness.

However, the clearest and most definitive example is the establishment of the Google Ethics Committee. On 26 May 2019, Google announced that it would launch an external advisory board responsible for AI development, the Advanced Technology External Advisory Council (ATEAC). Following a petition signed by 2556 Google workers demanding the removal of one of the company's board members, the anti-LGBT advocate Kay Coles James, the advisory body was disbanded about a week after its announcement. However instead of restructuring the committee and seeking representativeness for all its stakeholders, the tech giant chose to reject the committee altogether as an initiative. In the wake of this decision, it is very hard to believe in this company's sincerity and transparency. Instead, what it really shows is the arbitrariness and instrumentality of its business ethics management. The combination of opacity and power is today a threat to democracy.

The instrumentalization of ethics by business has led many authors to speak of the relationship between ethics and regulation as a zero-sum game. The role played by public commitments, the incorporation of ethical professionals, ethics councils or committees, training and capacity building of managers and employees, sustainability reporting, international guidelines, and so on. are paper tigers whose superficiality and ineffectiveness are quite evident. Furthermore, there is the *business of ethics*, the increasingly common presence of specialized consultancies and law firms that develop practices and tools for the ethical management of Artificial Intelligence and its companies and technologies. "The Partnership on AI" is a non-profit organization set up to study and formulate best practices. It was created by Amazon, Google, Facebook, Deepmind, Microsoft and IBM. A clear example of the fox protecting the chickens. Another example is provided by the range of international guidelines which developers must adhere to, only a couple of which actually specify measures to be applied (Hagendorff 2020; Calo 2017). The basic question is: do business ethics serve any purpose other than to conceal responsibility for a company's impact? In response to these doubts, would it not be better to forgo moral resources and

instead try to create legal resources? (García-Marzá 2022).

The question inherently assumes that the concept of ethics is being used superficially, instrumentalizing moral language; as if moral resources such as credit and trust were dispensable and had a merely cosmetic purpose. The need for and purpose of a business ethic becomes clear when we realize we are dealing with a mechanism of civil society within the democratic system, forming part of the effort to achieve a common will and use it for collective action. As part of this function, the responsibility of technology corporations becomes apparent if we start from a concept of the company as a socioeconomic institution whose legitimacy depends on the possibility of a free and voluntary agreement between *all the* interest groups involved in its activity (García-Marzá 2011; Green 2021a). Corporate responsibility does not depend on a unilateral calculation of results made by boards of directors and their chairpersons, but the moral validity of the contribution these make depends on the acceptance of these results by *all* the parties involved or their representatives, in a free and open dialogue conducted under equal conditions of participation (García-Marzá 2011).

The idea of dialogue and a subsequent agreement thus become a *normative concept*, a horizon of meaning from which to construct reasons for assessing the legitimacy of a company or any other type of institution, whether public, private, mixed, and so on. The distance to be covered until reaching the moral horizon of a possible agreement between all those involved under equal conditions finds its *empirical basis* in the trust or distrust that institutions deserve (García-Marzá 2012). When we speak of self-regulation, of self-obligation, we are not referring to managers or owners, but to the company as a socioeconomic institution. It is in this transition from a company understood as an instrument at the service of shareholder value, to a company capable of including all its stakeholders in the generation of value and in the monitoring and control of its business strategy (stakeholder governance), where the credibility of the company, its legitimacy and, therefore, the trust it deserves, are at stake. It is in the latter case that the theory of democracy and business ethics intersect, as a reconstruction and application of the ethical foundations underlying trust in business.

Within this dimension of deliberation and possible agreement, ethics becomes a management system complementary to law: while in law it relies on external coercion; in ethics it is the inner conviction that we should act in this particular way. There is no choice between legal regulation (state or European) and ethical regulation (business culture). Rather, we must be aware of the different logics of law and ethics in order to speak of a complementarity between the coercion of law and ethical commitment. While legitimate law already presupposes ethics, ethics is not limited to self-regulation, but its basic

contribution lies in the strength of moral values, in a culture predisposed to correct behaviour, to the ability to act in cooperation with others, always rowing in the same direction. Without this institutional character, without this way of thinking and being that we call “ethics”, law is impotent, and even more so in today’s global and technologically complex context.

Business activity responds to a series of expectations generated by the internal and external groups involved in or affected by it. If society, in the form of the public opinion created through these different interest groups, perceives that the company responds and fulfils these expectations with the right approach, then it grants the necessary trust and gives sufficient credibility to achieve and guarantee, for example, a good working environment or the acceptance of its products or services. When we speak of a moral contract, we refer to this *reciprocal play of expectations*. It is none other than the moral knowledge that builds these bonds of trustworthiness that concerns ethics, in order to make them part of an institutional design capable of generating trust.

However, taking this step requires more than a declaration of good intentions on the part of the practitioner and, in our case, the organization. On the basis of this principle of accountability, any form of information management that aims to account for its impact must, in each specific situation, involve *making public* the efforts it has made. It is not only a question of a willingness to be honest, but that this willingness should acquire the status of a *public commitment*, in the double sense of transparency and participation. Institutional design theories work with this idea by referring the capacity to produce trust to this “discursive justification potential” (Goodin 1998).

In other words, to address this public justification, the goodwill of a designated manager is not enough; we must have processes and organizational structures in place that enable and enhance ethical and accountable management. Public engagement requires both openness and accessible information, yet also the possibility of participation, in other words, the agreement of the parties involved in the organization’s activity. To anchor these ethical foundations of communication, we need both *transparency* in terms of information, and the *participation* of different stakeholders (García-Marzá 2012). It is now time to find an answer that demonstrates the difficulty, but not the impossibility of a positive response to the question: is it possible to think of the existence of an institutional design for the major technology companies that develop their resources and capacities to empower subjects, while defining institutionally ordered spaces, where different stakeholders or representatives can deliberate and participate in freedom and equality, without any fear of the consequences?

12.3 Corporate Public Sphere: At the Core of Big Tech

Democratic citizenship includes *corporate citizenship*, in the sense of a company's employees' sense of belonging to the organization, yet also referring to the company itself as an organization that is part of the democratic system. It is not just a question of employee participation (Brinkmann et al. 2022), but rather of all those involved in and affected by the activities of technology companies. Indeed, billions of people forming part of an institution are engaging in participation, however minimal it may be. We are talking about the participation of all stakeholders in the institutions of civil society, in this case the technology companies that build the technological and cultural resources underlying algorithmic democracy.

Again, it is not a matter of reproducing the demands of political participation in different institutions repeating the systems of election and decision-making by majority rule, in all kinds of social practices. The demand for participation must always be mediated with the social sense and function, the social good that institutions provide, and from which they draw their meaning. We are not saying that the responsibility of institutions begins and ends with their contribution to political power, but rather that the sense and purpose of the institution establishes the logic and meaning of this participation (Cortina et al. 2008).

The participation we seek has to be able to generate trust and this does not happen without an institutional design that sees the organizations themselves as *institutional actors* and what we are interested in is their contribution to democracy as institutions. Without this perspective, it is difficult, if not impossible, to account for the potential for social transformation encapsulated by Big Tech. Instead, what needs to be done is to design spaces for participation within institutions that guarantee the minimum conditions for free and equal communication between all the interest groups and all the parties involved: spaces for participation, dialogue and the search for inclusive agreements in the face of existing problems, but also spaces for deliberation, contestation and criticism, training, collaboration, and so on (Cornwall and Coelho 2007). Public spheres within these institutions through which to establish the necessary claim to legitimacy which underlies the trust they deserve. If legitimacy depends on the possible agreement of all those involved and affected by the institution, it is necessary to create spaces of trust where there is free and equal participation, and where there is deliberation, monitoring and control of the generalized interests at stake (García-Marzá 2019). Trust-building in turn requires spaces where trust can be assumed. Hence, the basic objective of the ethical perspective

of institutional design is the construction of an ethical infrastructure that includes these spaces of trust and participation in each of these institutions, as well as communication between them. Our question could be rephrased as follows: can spaces of participation be defined as sites where individuals' moral capacities combine in order to undertake a responsible commitment (Offe 2004)? Any possible response requires an institutionalist perspective capable of providing ethical principles for the design and redesign of institutions that help us to think of new and better political, social and economic institutions, that is, ones closer to the idea of justice (García-Marzá 2011; Schultz and Seele 2022).

The traditional mechanisms for ascertaining an organization's public commitment to ethical management, for example, have been organizational codes of ethics and, more recently, social responsibility reports or sustainability reports. However, as we will see below, we need to integrate these into a broader organizational design that includes, in line with the ideas outlined above, both transparency and participation. The first step in building trust is to draw up and publish a *Code of Ethics and Conduct*. These are formal documents containing an explicit statement of the values that should guide the conduct of employees and managers, so that they foster good practices and define an organization's character and personality (Schwartz 2004; García-Marzá 2017). Their function is two-fold: internally, to formalize the values and decision-making criteria that define the organization's specific culture and the behaviour expected of its employees; and externally, to manage the organization's reputation. In other words, they not only present us with the values that define an organization's character or ethics, but also the commitments it is willing to assume in order to achieve this common will and the conduct necessary for its realization (García-Marzá 2022).

The importance of these corporate codes grows with the global nature of many of the problems facing organizations today, whether they are companies, administrations or civic and solidarity organizations. In the face of these global and multicultural challenges, law, which is always limited to state borders, can do little, nor can tradition itself, which depends on each particular culture and religion. Ethical codes have no such state limitation. They present the organization's own values, commitments, and conduct, wherever it operates, thereby they demonstrate its specific character: its specific way of being and doing. Essentially, an ethical commitment must be in line with the moral values expressed in international guidelines and agreements, such as the UN's Global Compact or the Sustainable Development Goals (Stohl et al. 2009).

Nevertheless, the stated values' lack of specificity and the frequent non-compliance with the published commitments and regulated forms of behaviour

has cast doubt upon, and generated mistrust about these public commitments. The lack of participation from employees and other groups in developing them has undermined their effectiveness, and the frequent fraud and malpractice in their instrumentalization—as has been discussed—have undermined the credibility of these self-regulatory mechanisms in the eyes of the public (Metzinger 2019). For many organizations, ethical codes are as far as their public engagement goes, as if it were enough to merely make a declaration of goodwill in order to be credible. They forget that codes of ethics are the first step in building trust, not the last, and let alone the only one.

They have been supported by social responsibility reports or sustainability reports as standardized procedures, e.g., emulating the *Global Reporting Initiative*, or the Sustainable Development Goals for the presentation of the economic, social and ecological results generated. Essentially, they provide information on the impact generated by the organization. While transparency is justified in these reports through reports on management policies and strategies, as well as results in the form of indicators; nevertheless, they fail to provide confidence in these companies' ethics and compliance programmes, in short, they fail to build trust.

Again, the reason lies in the disconnection between transparency and participation, i.e., in the verticality of information. While they have achieved a common form of reporting compliance through the code of ethics, the required materiality reports are not sufficient to account for stakeholder engagement, especially external stakeholders. Their preparation by communication or corporate social responsibility departments or, worse still, by specialized consultancy firms, who claim to verify the information provided, fails to achieve credibility. Another question to be addressed in more detail below, is whether these reports are understood as arguments forming part of agreements made by the corporate ethics committee.

The proposal advocated here does not renounce these two steps, which are absolutely necessary to generate confidence in the information offered. Instead, it proposes to integrate a need for transparency with a need for participation, so that these two instruments for managing ethics at the company are supported by the participation of all the interest groups involved in and affected by organizational activity. The first step in building trust is therefore to include compliance mechanisms in the code of ethics itself, as part of the code and not as separate parts. In this sense, we are talking about a renewal, indeed a *new generation of codes of ethics*, precisely because they incorporate participation into the definition of corporate commitments, and also into the monitoring and control of their fulfilment.

This new generation of ethics codes moves beyond current ethics and compliance programmes, which are always limited to the legal sphere and concerned only with criminal risks (Peterson 2013; Weber and Wasieleski 2013). They fail to take into account *moral hazard*—the possibility that an organization does not behave in accordance with expectations of it—and that this can have still worse repercussions. Credibility and trust are at stake here. So, these two remaining components aim to bring the communication of social responsibility closer in line with participation.

This is why there is a need to think, debate and jointly design spaces of trust that develop institutions' capacity to empower subjects within institutionally ordered, flexible, open spaces, where they can deploy their capacities for understanding and agreement. Digital technologies can help with the creation and maintenance of these spaces for presence, but not replace them with virtual realities whose ultimate aim is to control and commercialize our opinion and our will, and thereby conceal the injustice that exists. However, this form of participation must be specifically established at companies with millions of shareholders, billions of customers and well-known products capable of operating across all borders and with opaque management. How can we even consider the participation of all the parties involved as a possibility, however minimal that may be? Examples of these spaces of trust are Corporate Ethics Committees (CECs), which the authors have been working on for decades.

A CEC is conceived as a space for participation and dialogue between the different stakeholders within an organization, and it is in charge of monitoring and controlling the company's ethics and compliance programme, as well as promoting ethics and social responsibility. Its function is three-fold: to *advise* on issues concerning the interpretation and application of a code of ethics; to *resolve* suggestions, alerts and complaints made via the ethics line; and to *promote* information and training for employees and managers in the ethics and compliance programme. Its members must reflect the main interest groups making up the organization, and remain constantly aware that the CEC's aim is to assist in managing and ensuring compliance with the values and standards expressed in the Code of Ethics and Conduct. Trust in the committee will depend on the trust that its members are capable of generating, so consideration must be given to the transparency and participation mechanisms (García-Marzá 2005).

For example, reports and other corporate information must be prepared by the company's professionals, but they must always pass through the filter provided by the CEC, a space which must also include external stakeholders. However, it is not a matter of repeating the systems of election and decision-making by majority rule at the heart of Big Tech and the other institutions and

organizations making up the current socio-technical context. This is neither possible nor desirable, as is demonstrated, for example, by the failed experiences of industrial democracy, and even by the relationship between corruption and political participation. We are not talking about civil society institutions as a "secondary circuit for politics", but about protagonism and civil society, as well as the companies' assumption of responsibility for the public sphere.

Assistance is provided by one of the oldest and most significant mechanisms of democratic participation: selection by lottery (Keane 2009). If we are able to complement this demand for participation with new digital technologies, it becomes possible to think together about desirable and possible solutions. We will focus on two innovative examples: selection by lottery, and the ethics line as a channel for communication. These are two key elements of what we can term *an ethical infrastructure* for companies, an institutional design capable of redressing opacity and generating trust in the communication and management of impact and, at the same time, of building a retaining wall against technological colonialism. We will now turn to examine some principles that may help in this design:

1. *Explicability principle*: the core values and commitments the company intends to adopt must be clearly and explicitly defined, i.e., publicly stated. This means defining a code of ethics and compliance that establishes both the rules of good governance and expected behaviour, as well as the measures to ensure their control. It is this that provides the ethical framework for corporate action and the focus for drafting reports and statements.
2. *Principle of inclusivity*: all stakeholders should be identified as exhaustively as possible, and any interests at stake and potential conflicts should be defined. Each of these groups will draw up a list of potential participants for the CEC and other participatory activities. Minimum numbers will be established to determine the inclusion of all those who wish to be included.
3. *Principle of participation*: to select stakeholder representatives, two members representing each stakeholder group will be chosen by lottery, after registration in the corporate citizenship list. They will form part of the ethics committee for a maximum period of four years.
4. *Principle of equality*: the committee addresses only universalizable interests, the basic rights underpinning decision-making in the company, and it always

acts on the basis of the agreement of all members.

12.4 Spheres of Participation and Communication

The participation sought through CECs cannot be limited to a small committee; although, it gives a presence and voice to internal and external stakeholders it is no substitute for the voice of everyone. We must therefore establish channels of communication that enable the participation of all those who wish to express opinions or concerns, while always remaining focused on the fulfilment of the ethical commitments adopted. Communication should not be limited to reporting malpractice. Instead, and above all, an ethical culture entails the involvement of employees in alerting, preventing and detecting situations and conflicts of interest that could lead to breaches harming the company's reputation (Lee and Fargher 2013; Calvo 2018a, b). The same is true for external groups, users, customers, suppliers, and so on. The aim is to move towards an ethical culture where the prevention and detection of malpractice become a shared responsibility.

This role is performed by the *ethics line*, which is understood as a participation channel expressing suggestions, alerts and complaints. Using any format (post, email, web form, etc.), its aim is always to communicate breaches of the code of ethics or suggestions for improvement. Communication should, of course, always be confidential, but never anonymous (Elka and Carey 2016). Protocols should be put in place to ensure confidentiality and avoid any possible retaliation. Thus, suggestions, alerts and complaints should be addressed to the ethics committee and not solely to the ethics and compliance officer. Essentially, we are concerned with different yet sequential steps for generating trust within a company. These steps are not independent, they constitute a system in the sense that they refer to each other when we think of managing trust, as that is the final outcome of this interrelationship. However, the existence and operation of this ethical infrastructure subsequently has to be verified externally, which is the purpose of ethical auditing. This is undertaken in the form of the explicability report.

A design for an ethics line like the one proposed here can be used as a key element in the implementation and development of spaces for participation and trust across a range of contexts of activity. This is partially due to the fact that it is a tool that can help: provide a diagnosis of the general rate of compliance with ethical standards across a sector, or within an institution or organization; detect or prevent malpractice; receive suggestions and proposals for improving on the content and application of the code of ethics, the rules of conduct and the

procedures involved; detect conflicts, challenges and opportunities; assess the gap between what is said and what is done and so on.

Today, however, the process of digital transformation and its related techniques and technologies offer new ways and possibilities for managing and developing ethics lines through the application of *blockchain* technology, machine learning and *data mining* techniques. Applying these developments makes it possible to improve the prospects of applying and implementing the proposed principles of explicability, inclusivity, participation and equality. These are two examples of how digital technologies can help to restore participation and deliberation within institutions, however large they may be.

12.4.1 Ethics Lines and Blockchain

The *blockchain* technology described in Chaps. 1 and 4 firstly offers a substantial improvement in terms of the reliability, traceability, uniqueness, accessibility, verifiability and confidentiality of the information collected by the ethics line. It also gives increased protection for *whistleblowers*, while guaranteeing an appropriate and responsible use of the information supplied. In addition, *blockchain* technology offers the possibility of turning the ethics line into an intelligent system that complements the human processes of controlling and auditing the system, as well as the irreplaceable task of the *whistleblower*. A *blockchain ethics line is*:

– *Unalterable:* it permits data to be generated, stored and organized in a codified way, and without any need for a central entity to record the submission of suggestions, good practices, complaints and alerts of ethical non-compliance. The great advantage of all this is that records cannot be deleted or modified, which maintains the information's integrity, uniqueness and traceability.
– *Auditable:* it ensures the contrasting, verification and tracking of information concerning suggestions, good practices, complaints and alerts and/or reports of ethical non-compliance of an institution, organization or company and its professionals through AI algorithms.
– *Participatory:* the internal and external stakeholders of the institution, organization or company can be informed about the company's degree of ethical compliance via a dashboard that presents the data records and metadata collected by the ethics line. The inclusion and critical participation of internal and external stakeholders is thereby encouraged and promoted in trusted spaces and as part of continuous improvement process linked to ethics.
– *Anonymized:* the identity of system users is safeguarded through the

encryption of their personal data, and *anonymization of* information, thereby enhancing the protection for *whistleblowers* even though they must give their names. The system's algorithms take care of the safekeeping and management of *whistleblowers*' sensitive data to avoid any possible retaliation.
– *Monitorable:* it permits the inclusion and use of *ethical cyber-blowers* in the system. In other words, AI algorithms that can scrutinize the institution, organization or company to suggest improvements to the system and/or alert the ethics and compliance officer and members of the ethics committee to possible anomalies or cases of ethical non-compliance.

12.4.2 Ethics Lines and Data Mining

Data mining techniques make it easier to see patterns that help those involved or affected to identify possible behavioural anomalies in institutions, organizations and companies and their professionals, as well as to act on them in order to correct, eradicate or mitigate these weaknesses. Any potential anomalies are analysed by the line managers so they can be acted on or discarded.

Data mining is a set of techniques used for characterization, discrimination, association, classification, prediction, clustering, *outlier* analysis or evolution processes involving data and it constitutes a further level within the *Knowledge Discovery in Databases* (KDD) process. In this sense, its ability to produce relevant information through the exploration and analysis of large amounts of data has made it a core element of the process because it can reveal patterns within a dataset. In this respect, two techniques stand out as relevant for the practical development of *ethics lines*: *clustering* and *association rules*.

– *Clustering* is a subset of *data mining*, and involves different unsupervised learning techniques whose main aim is to find patterns that allow for the grouping or segmenting of similar elements or objects, such as a certain type of behaviour, within a set of observations using different types of clustering algorithms (Amat-Rodrigo 2018; Celebi 2015, 2016; Nasraoui and Ben N'Cir 2019).
– *Association rules*, another subset of data mining, is a technique that uses machine learning algorithms to identify robust rules in large databases based on measures of interest (Piatetsky-Shapiro 1990, 1991); in other words, "(...) relationships within a set of transactions, namely items or attributes that tend to occur together" (Amat-Rodrigo, 2018).

In short, the mature, critical and prudent application of *data mining* techniques offers significant possibilities for the current development of ethical and other lines:

– Revising information and diagnosing a situation regarding the moral behaviour of an institution, organization or company and its professionals, as well as detecting anomalies, potentialities and/or weaknesses from an ethical perspective.
– Extracting relevant information to improve the understanding of the behaviour, whether right or wrong, of an institution, organization or company and its professionals. For example, personalizing training programmes for a department, area or work group; developing a necessary focus on a specific aspect in order to strengthen or eradicate it; detecting flaws in the system itself; addressing the lack of a specific value in the code of ethics, and so on.
– Identifying indications of possible critical and/or post-conventional behaviour through the analysis of *outliers*; this permits influence to be exerted over the practical development of an institution, organization and company and its professionals and, to thereby conduct specific and personalized actions to include or strengthen them.

However, there are still many questions to be answered and constraints to be overcome for its proper implementation. These include how to overcome the challenge posed by the energy inefficiency of *blockchain* systems; how to build trust in AI-based processes for the analysis and interpretation of outliers in order to detect fraud, behavioural anomalies and, above all, post-conventional criticism within the information generated by the institution, organization or company; how to address the variability of the items or attributes used in an AI behavioural assessment in accordance with the contextual reality of an institution, organization or company that designs, applies and/or uses it; and finally, how to design and implement effective, accessible and revisable mechanisms in order to implement the principle of explicability in practice, and human control within the design, application and use of AI (Calvo and Egea-Moreno 2021).

12.4.3 Explicability Report

The explicability report is a formal document intended to promote self-knowledge, continuous improvement and communication wherever an institution, organization or company publishes its degree of compliance with the principles and requirements of the European ethical guidelines *(High-level Expert Group* on *Artificial Intelligence* 2019); the values, forms of behaviour and commitments established by an institution, organization or company's code of ethics; and, finally, the economic, social and environmental impacts generated by the design, application and use of AI as part of its business activity. This

report is therefore included in and forms part of the Social Responsibility Report in order to give an exhaustive, intelligible, sincere and concise account of the company's aims and mechanisms, as well as the economic, social and environmental impact resulting from the use of AI. It is based on (general and specific) measurable, comparable and auditable indicators such as those given below.

The explicability report thus provides a specific complement to the Social Responsibility Report for institutions, organizations and companies that design, develop, apply and/or use AI algorithms**.** It is also linked to the ethics audit as an indispensable instrument to verify the information collected, generated and published by the ethics line, as well as critically analysing the level of intelligibility, veracity, truth and justice achieved by the institution, organization or company that designs, applies or uses AI (García-Marzá 2017). Finally, the explicability report becomes a fundamental tool that can be used by an ethics committee to improve its assessment of an institution, organization or company's activity, its approach to preventing and resolving the ethical conflicts underlying the design, application and use of AI, and, finally, its contribution to constructing a business culture and a character that satisfies the desires and demands of the twenty-first-century hyperdigital society.

12.5 Spaces for Trust: Recovering Participation

Together, these two clear examples of the integration of technological advances and their algorithms into democratic theory, provide an illustration of a possible non-political structuring of civil society. The design of spaces for participation by all involved and affected parties within institutions, whether they are platforms, large corporations, universities, research, technology and innovation centres, or whatever, can be achieved in corporate public spheres under the principles of inclusion, equal participation and the consensual search for agreements on interests capable of generalization. This establishes a brake on the current algorithmic colonization and the rapid replacement of public policy decisions by algorithms.

Over the course of this book, we have demonstrated how aggregative democracies do not constitute a defensive measure against the advances of algorithmic democracy, instead proving easy prey. Their conception of politics as an addition and subtraction of particular interests, pre-legitimitized by the mere fact of the existence of such interests, prevents any search for anything shared. Instead it reinforces an autistic individualism whose desires and preferences are ultimately the product of the major technological companies,

which are hegemonic yet remain invisible within concepts of democracy. In this way, the reduction of democracy to representation and, in turn, merely to elections, promotes technopopulism and, essentially, algorithmic colonization. It might be suggested that we are exaggerating, but all the current indexes measuring democratic quality, ranging from “The Economist” to “Freedom House”, signal a worldwide decline in democracy, as well as increasing authoritarianism. The new global scenario presents us with an increase in anti-democratic alliances, which pose a major challenge to our democracies.

Nevertheless, participatory deliberative democracies’ two-way structures allow the introduction of civil society and its institutions as a sphere of power, and therefore of responsibility. Here, the democratic principles of inclusion, equality and participation must also be applied. As a result, civil society and its own coordinating mechanisms become a powerful brake for the advances of algorithmic colonization. In this final chapter, we have examined the institutional design of the institutions making civil society and, in particular, the major technology companies, who are the true protagonists of the construction of the current socio-cultural context. They are responsible for the disastrous consequences democracy currently faces, as has been discussed over the course of this book. The intention underpinning the design of spaces for participation—spaces for trust and reciprocity within institutions themselves—can be achieved through the proposal of using applied ethics to structure civil society as set out above. This is the potential provided by the moral knowledge that underpins our democracies; essentially the relationship between ethics and democracy.

We turn, then, to our main proposal: the creation of spaces for participation, trust and proximity as part of an ethical infrastructure for big technology as a response to the principles introduced for institutional design. The concluding discussion, in the final two sections, has referred to the implementation of this proposal through a consideration of two examples: the design and creation of an ethics line that allows the direct participation of all parties involved and exploits blockchain technologies; and, finally, the design of an explicability report, which always allows for the possibility of transparency and participation. Together, these two clear examples of the integration of technological advances and their algorithms into democratic theory provide an illustration of a possible non-partisan structuring of civil society. Participation is unquestionably the most direct and secure way to restore confidence in our institutions.

As the example of Google's ethics committee shows, the mechanisms for choosing the components of these public spheres do not depend on the whim of managers or vested interests. In these cases, trust disappears. The mechanisms of representation can range from exemplary democracy to trust in a lottery. We are

not dealing with “mini-publics”, as in these spheres the social good that each institution provides (education, health, research, and so on.) is dealt with in these areas. Their functions range from monitoring and controlling the public commitments acquired by the corporation to managing the communication channels for alerts, suggestions and complaints (Walzer 2002). They operate by creating networks of trust and *antennas* capable of creating public opinion concerning the general interests these corporate actors are capable of providing or destroying (García-Marzá 2017).

We are not talking about extending voting and majority rule to all institutions and platforms, for example. This *democratism*— this political colonization— would be neither possible nor desirable.

Two-way democracy is the main contribution of deliberative democracy, where civil society, based on the primacy of the communicative action that characterizes it, has the possibility of creating spheres of trust where the search for common interests, often in conflict, can be satisfied through deliberation and possible agreement always based on the different spheres of justice constituting it (Walzer 2002). As face-to-face spaces for participation and deliberation, corporate public spheres represent the surest way to break through echo chambers and bubble filters. In short, they represent the design spaces for presence at a *mesodeliberative level.*

Participation is unquestionably the most direct and secure way to restore confidence in our institutions. To conclude with the words of Offe: “From a social perspective, intermediary associations can ensure that the capacity for commitment to the norms considered as valid is generalized within the civic society, so that none of those involved need fear that their own regulated behavior will go unanswered by others or that it will be exploited (…) Designing spaces of participation where the moral capacities of individuals converge to assume a responsible commitment” (Offe 2019).

Bibliography

Amat-Rodrigo, J. 2018. Reglas de asociación y algoritmo Apriori con R. Cienciadatos, *Github*. https://www.cienciadedatos.net/documentos/43_reglas_de_asociacion#Bibliograf%C3%ADa.

Bietti, E. 2019. From Ethics Washing to Ethics Bashing A View on Tech Ethics from Within Moral Philosophy. In Proceedings of ACM FAT* Conference (FAT* 2019). ACM, New York, NY, USA.

———. 2020. From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics. *Journal of Social Computing* 2 (3): 266–283. https://doi.org/10.23919/JSC.2021.0031.
[Crossref]

Brinkmann, U., H. Heiland, and M. Seeliger. 2022. Corporate Public Spheres Between Refeudalization and Revitalization. *Theory, Culture & Society* 39 (4): 75–90.

Calo, R. 2017. Artificial Intelligence Policy: A Primer and Roadmap. *UCDL Rev* 51: 399.

Calvo, P. 2018a. Whistleblowing y datos masivos: monitorización y cumplimiento de la ética y la responsabilidad social. *El Profesional de la Información* 27 (1): 173–184.

———. 2018b. *The Cordial Economy – Ethics, Recognition and Reciprocity*. Cham: Springer.

Calvo, P., and Rebeca Egea-Moreno. 2021. Ethics Lines and Machine Learning: A Design and Simulation of an Association Rules Algorithm for Exploiting the Data. *Journal of Computer Communications* 9 (12): 17–37.

Celebi, M. Emre, ed. 2015. *Partitional Clustering Algorithms*. Springer: Cham.

———, ed. 2016. *Unsupervised Learning Algorithms*. Springer: Cham.

Cornwall, Andrea, and Vera Schatten Coelho, eds. 2007. *Spaces for Change?: The Politics of Citizen Participation in New Democratic Arenas*. London: Zed Books.

Cortina, Adela. 1993. *Ética aplicada y democracia radical*. Madrid: Tecnos.

Cortina, A. 2007. Development Ethics: A Road to Peace. *Kellogg Institute for International Studies Working Paper*, (339).

———. 2021. Ética cosmopolita. *Una apuesta por la cordura en tiempos de pandemia. Editorial Planeta, 1ra edición, España*.

Cortina, A., J. Conill, D. García-Marzá, and (Eds.). 2008. *Public Reason and Applied Ethics: The Ways of Practical Reason in a Pluralist Society*. Ashgate Publishing, Ltd.

Delacroix, S., and B. Wagner. 2021. Constructing a Mutually Supportive Interface Between Ethics and Regulation. *Computer Law & Security Review* 40: 105520.

Elka, Johansson, and Peter Carey. 2016. Detecting Fraud: The Role of the Anonymous Reporting Channel. *Journal of Business Ethics* 139: 391–409. https://doi.org/10.1007/s10551-015-2673-6. [Crossref]

Fjeld, J., N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *Berkman Klein Center Research Publication* (2020-1).

Fung, A. 2003. Survey Article: Recipes for Public Spheres: Eight Institutional Design Choices and Their Consequences. *Journal of Political Philosophy* 11 (3): 338–367.

Garcia-Marzá, D. 2005. Trust and Dialogue: Theoretical Approaches to Ethics Auditing. *Journal of Business Ethics* 57: 209–219.

———. 2008. Sociedad civil: una concepción radical. *Recerca: revista de pensament i anàlisi*: 27–46.

———. 2011. *Ética empresarial: del diálogo a la confianza*. Madrid: Trotta.

———. 2012. Kant's Principle of Publicity. *Kant-Studien*. *Philosophische Zeitschrift der Kant-Gesellsschaft* 103: 96–113.

———. 2013. La dimensión ética del diseño institucional. In *Ética y gobernanza, Un cosmopolitismo para el siglo XXI*, ed. E. Gonzàlez Esteban. Grenada: Comares.

———. 2017. From ethical codes to ethical auditing: An ethical infrastructure for social responsibility communication. *Profesional de la información* 26 (2): 268–276.

———. 2019. Deliberative Democracy: An Upgrade Proposal. In *Routledge Handbook of Contemporary European Movements*, ed. Flescher Fominaya and Ramón Feenstra. London: Routledge.

———. 2022. El valor de las instituciones en la sociedad digital: algunas consideraciones desde la Teoría de la Justicia. In *Actualidad de John Rawls en el siglo XXI*, 163–178. Granada: Comares.

González, E. 2002. Defining a Post-conventional Corporate Moral Responsibility. *Journal of business ethics* 39 (1): 101–108.

Goodin, R.E., ed. 1998. *The Theory of Institutional Design*. Cambridge University Press.

Green, B. 2021a. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing* 2 (3): 249–265.

———. 2021b. The Contestation of Tech Ethics: A Sociotechnical Approach to Ethics and Technology in Action. *arXiv preprint arXiv:2106.01784*.

Haas, L., S. Gießler, and V. Thiel. 2020. In the Realm of Paper Tigers-Exploring the Failings of AI Ethics Guidelines. Accessed 29 June 2021.

Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30 (1): 99–120.

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission. Available at https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

Hyatt, J.C. 2006. You're the New Ethics Officer, Now What? *Business Ethics: The Magazine of Corporate Responsibility* 20 (1): 30–34.

Keane, J. 2009. *The Life and Death of Democracy*. Simon and Schuster.

Larson, E.J. 2021. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Harvard University Press.

Laufer, William S. 2003. Social Accountability and Corporate Greenwashing. *Journal of Business Ethics* 43 (3): 253–261.

Lee, Gladys, and Neil Fargher. 2013. Companies' Use of Whistle-Blowing to Detect Fraud: An Examination of Corporate Whistle-Blowing Policies. *Journal of Business Ethics* 114: 83–295.

Martin, K.E. 2020. Ethical Issues in the Big Data Industry. In *Strategic Information Management*, 450–471. Routledge.

Metcalf, J., and E. Moss. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalisation of Ethics. *Social Research: An International Quarterly* 86 (2): 449–476.

Metzinger, T. (2019). Ethics Washing Made in Europe. *Der Tagesspiegel*, 8.

Nasraoui, Olfa, and Chiheb-Eddine Ben N'Cir, eds. 2019. *Clustering Methods for Big Data Analytics. Techniques, Toolboxes and Applications*. Cham: Springer.

Nemitz, P. 2018. Constitutional Democracy and Technology in the Age of Artificial Intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133).

Offe, C. 2004. Political Institutions and Social Power: Some Conceptual Explorations. In *Gesellschaft mit beschränkter Hoffnung*, ed. P. Stykow and J. Beyer. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-80467-9_3.
[Crossref]

———. 2019. Democratic Institutions and Moral Resources. *Liberale Demokratie und soziale Macht: Demokratietheoretische Studien*, p. 139–165.

Peterson, Evan A. 2013. Compliance and Ethics Program: Competitive Advantage Through the Law. *Journal of Management & Governance* 17 (4): 1027–1045. https://doi.org/10.1007/s10997-012-9212-y.
[Crossref]

Piatetsky-Shapiro, G. 1990. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Magazine* 11: 68–70. https://doi.org/10.1609/aimag.v11i4.873.
[Crossref]

———. 1991. Discovery, Analysis, and Presentation of Strong Rules. In *Knowledge Discovery in Databases*, ed. G. Piatetsky-Shapiro and W.J. Frawley, 229–246. Cambridge: MIT Press.

Powell, W.W., and P.J. Dimaggio. 1991. *The New Institutionalism in Organizational Analysis*. Chicago: University Press.

Robbins, Scott, and Adam Henschke. 2017. Designing for Democracy: Bulk Data and Authoritarianism. *Surveillance & Society* 15 (3/4): 582–589.

Sætra, H.S., M. Coeckelbergh, and J. Danaher. 2022. The AI Ethicist's Dilemma: Fighting Big Tech by Supporting Big Tech. *AI and Ethics* 2 (1): 15–27.

Schultz, M.D., and P. Seele. 2022. Towards AI Ethics' Institutionalization: Knowledge Bridges from Business Ethics to Advance Organisational AI Ethics. *AI and Ethics*: 1–13.

Schwartz, M.S. 2004. Effective Corporate Codes of Ethics: Perceptions of Code Users. *Journal of Business Ethics* 55 (4): 321–341.

Sorsa, Ville-Pekka. 2008. How to Explain Socially Responsible Corporate Actions Institutionally: Theoretical and Methodological Critique. *Electronic Journal of Business Ethics and Organization Studies* 13 (1).

Stohl, Cynthia, Michael Stohl, and Lucy Popova. 2009. A New Generation of Corporate Codes of Ethics.

Journal of Business Ethics 90 (4): 607–622.

Thao, Phan, Jake Goldenfein, Monique Mann, and Declan Kuch. 2021. Economies of Virtue: The Circulation of ‘Ethics’ in Big Tech, Science as Culture.

Treviño, Linda K., et al. 1999. Managing Ethics and Legal Compliance: What Works and What Hurts. *California Management Review* 41 (2): 131–151.

Veale, M. 2020. A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence. *European Journal of Risk Regulation* 11 (1): 1–10.

Walzer, M. 2002. Equality and Civil Society. *Alternative Conceptions of Civil Society*: 34–49.

Warren, Mark E. 1996. What Should We Expect from More Democracy? Radically Democratic Responses to Politics. *Political Theory* 24 (2): 241–270.

Weber, James, and David M. Wasieleski. 2013. Corporate Ethics and Compliance Programs: A Report, Analysis and Critique. *Journal of Business Ethics* 112 (4): 609–626.

Index

A

Accelerating dynamics 15
Accessibility 68, 69, 71, 239
Accountability 11, 57, 72, 74, 109, 111, 118, 119, 158, 162, 182, 185, 196, 198, 216, 234
Agents of justice 211, 217, 229, 230
Aggregative democracy 15, 242
Algocracy 11, 12, 97
Algo-democracy 12
Algorithmic
  algorithmization of inequality 156
  algorithmization of injustice 156
  colonization ix, 6, 15, 155–159, 172, 183, 184, 242, 243
  construction 210
  culture 10, 11, 96, 175–179, 184, 224
  democracy ix, x, 4, 9–13, 47, 49, 51, 52, 96, 97, 140, 168, 171, 174, 175, 178, 179, 182, 187, 211, 216, 224, 234, 242
  governance 44–57, 92, 109, 111, 161
  ideology 170, 171, 173, 175
  morality 51
  neutrality 52, 156, 170–175
  objectivity 51
  paternalism 11
  politician 10, 105, 153
  reason 150, 170, 187
  representativeness 52
  transformation of morality 150–155
Algorithmization 30, 41, 43, 49, 90, 126, 154, 156, 174, 176
Algorithms 4, 25, 44, 64, 83, 104, 151, 169, 192, 209, 231
Algosphere 97, 98, 162
Alisa 29, 44–46, 55, 56, 106, 153
AlphaGo 116
Amoral 155, 157, 229
Analytics tools 89, 90
Antagonisms 6
Anthropomorphization 34

Apel, K.O. 12, 139, 149, 159, 170, 192, 196, 212, 213
Aporophobic 55, 56, 72, 111, 137, 160
Applied ethics 108, 138, 140, 194–197, 200–204, 209–216, 228, 243
Aproblematic 176, 177
Argumentative reasons 8
Aristotelianism 149
Artificial
  confidence 93
  ethics 28, 34–36, 53–57, 107, 108, 117
  hyperethics 157, 158
Artificial Intelligence (AI) 10, 25, 42, 62, 90, 103, 150, 169, 193, 223, 232
  AI-based tools 96
  algorithm 34–36, 52–54, 64, 90, 98, 152, 153, 157, 158, 160, 161, 240, 242
  applications 31, 114
  ethics committee 119
  Mayor 29, 44, 46, 47, 49, 50, 55, 106, 153
  models 98, 151
  moral agency 157
  moral conditionality 158
  moral meaning 157
  moral objectivity 158
  moral reality 158
  neural networks 27, 30, 108–111, 154
  public opinion 85, 87, 92
  public sphere 93–98
  social actors 94
Artificial Intelligence Act 118, 120
Artificialization 34, 92
Artificial moral intelligence (AMI) 108–111, 117
Artistic heritage 129, 130
Asepsis 175, 177
Asimov, I. 118
Assessments 52, 54, 73, 75, 106, 112, 120, 155, 170, 177, 181, 183, 194, 196, 202, 241, 242
Association roles 89
Auditability 74, 199
Augmented
  democracy 63–76
  man 117

participation 64, 65, 68–70, 74, 202
Authoritarianism 7–9, 186, 230, 243
Automated direct democracy 65
Automation
of inequality 177
of morality 148–163
Autonomous
cars 104
driving technology 105
subjects 3
Autonomy x, 7, 8, 10, 12, 14, 18, 35, 84, 97, 104, 109, 118, 137, 156, 158–160, 163, 169, 171, 172, 195–199, 201–203, 210, 212, 213, 216, 217, 224, 230
Axiological 112, 113, 120, 139, 149, 158, 163

B

Babbage, C. 26
Back-propagation algorithm 28
Beliefs x, 6, 10, 13, 97, 103, 104, 117, 150, 153, 154, 169, 173–176, 178, 180, 181, 184–186, 195, 202, 220
Benefits of AI 171
Bias 11, 31, 43, 45, 52, 54–56, 72, 104, 107, 111, 160
Big business 4
Big data 10, 12, 14, 30, 31, 43, 52, 72, 73, 88–92, 110, 126, 150, 161, 162, 168, 170, 180, 192–194, 198, 203
Big data technology 75
Big tech ix, 89, 98, 158, 163, 171, 230–238
Bioethics 154, 198
Bioethics committees 154
BlessU-2 152
Blind trust in algorithms 173
Blockchain
technology 64, 74, 75, 127, 128, 132, 133, 239, 243
Boden, M.A. 31, 114, 117
Boole, G. 26, 27
Bots 73, 88–92, 94–96, 133, 136
Brain
brain-computer interfaces 183
brain puncture 116
brain-society 182

C

Citizens 4, 33, 41, 82, 106, 150, 172, 192, 208, 230
Civic competence 8
Civic virtues 138, 148, 150
Civil
  disobedience 16
Civil Law Rules on Robotics 118
Civil society
  institutions 212, 214, 220, 221, 237
Climate change 63, 131, 186
Clustering 49, 89, 240, 241
Codes of ethics and conduct 75
Collective life x, 216
Collectivity 69, 71
Colonization of the lifeworld 89, 154–159
Common good 8, 43, 44, 46, 48, 52–54, 56, 75, 83, 84, 109, 148, 153
Commonplace 6
Communicative action 16, 18, 20, 155, 211–214, 217, 218, 244
Communicative power 221
Complexity
  of digital contexts 117
Conflicts of interest 10, 46, 238
Conill, J. 28, 35, 110, 126, 137, 139, 148, 149, 152, 159, 161, 180
Connected things 110
Connectionist model 27
Consensus 6, 9, 36, 70, 75, 82, 132, 155, 162, 196, 201, 218, 221
Consequence 6, 10, 27, 35, 36, 42, 43, 46, 54, 56, 57, 68, 71, 72, 74–76, 90, 91, 98, 107, 110, 115–117, 120, 128, 129, 135–139, 149, 155, 158, 160–162, 168–172, 175, 177, 187, 192, 194, 198–200, 202, 210, 223, 224, 228, 234, 243
Consistency 32, 90, 168, 175
Construct meaning 168
Conventionalism 160, 161
Cordial 139, 149, 160
Corporate citizenship 234, 238
Corporate public sphere 221, 234–238, 242, 244
Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) 151
Corruption 4, 5, 43, 46, 51, 65, 66, 75, 90, 105, 109, 186, 187, 215, 237

Cortina, A. 6, 7, 114, 116, 139, 148, 149, 159, 172, 196, 209, 210, 212, 213, 234
Credibility 4–6, 53, 82, 84, 93, 148, 155, 159, 168, 175, 195, 202, 208, 211, 216, 219, 221, 224, 228, 229, 233, 236, 237
Crisis 5, 33, 50, 62, 130, 131, 168
Criterion of rationality 139, 149, 157
Critical
  hermeneutics 19, 210, 213–215, 218
  judgement 110, 137
  perspective 7, 10, 12, 13, 17, 82, 97, 155, 170, 171, 173, 177, 193, 200, 208, 209, 211–213, 218, 229
Criticism 47, 93, 107, 119, 152, 184, 231, 235, 241
Crypto
  crypto-art 127–131
  cryptocurrency 128, 129, 131–133, 140
  cryptoeconomics 127–133
  cryptographic platform 128
  crypto-investors 129, 130
  cryptovirtual 132, 133
Cultural
  machines 176, 177
  processes 172
Cyber
  cyberdemocracy 11, 220
  cyberethical ecosystem 74, 75
  cyberethics 194
  cyber-physical ecosystem 35, 36, 63, 65, 89, 160
  cyberspace 93, 94, 96, 98, 127, 176, 180
  cyberutopia 171

D

Data 10, 26, 43, 62, 83, 104, 150, 169, 192, 229
Data ethics 194
Datafication
  of the real 169
Dataficablity 69
Data governance 196
Data Governance Act 120
Dataism 176
Data mining 49, 70, 89, 90, 92, 154, 203, 239–241

Datapositivism 170
Decision-making x, 6–8, 10–12, 19, 30–33, 36, 47, 53–55, 57, 63–65, 67, 70, 72, 83, 90, 96, 104, 105, 107–111, 114, 115, 118, 120, 126, 132, 151, 154, 156–158, 160, 170, 172, 173, 177, 179–181, 183, 194, 196, 198, 210, 215, 219, 222, 234, 235, 237, 238
Deep neural networks 28, 29, 31, 111, 114
Deliberation 10, 14, 16, 19, 35, 36, 44, 57, 70, 74, 83, 96, 154, 157, 159, 162, 169, 171, 179, 180, 196, 202, 204, 208, 210, 214, 220–222, 224, 233, 235, 239, 244
Deliberative democracy 13, 14, 17, 140, 159, 179, 209–211, 214–218, 223, 224, 243, 244
Deliberative politics 15, 17, 208
Democratic
  decisions 66, 82
  disaffection 168
  drift 4–20
  dynamism 13
  institutions 4, 63, 67, 75, 84, 179, 187
  legitimacy 5, 82, 84, 172
  models 7, 12, 43, 75
  oligopoly 12
  participation 4, 15, 16, 49, 238
  processes 10, 11, 15, 16, 35, 64, 65, 69, 90, 93, 95, 159, 172, 208
  revisionism 7–9, 13, 84, 172
  sphere 34–36, 90, 94, 120
  system x, 4, 5, 7, 8, 11, 12, 35, 43, 52–54, 56, 66–74, 84, 90, 98, 168, 179, 185, 200–204, 208, 209, 211, 213, 216, 222, 229, 230, 232, 234
Democratism 244
Deontological ethics 149
Deontology 118, 149
Design principles 66–72, 223
Det Syntetiske Parti 44, 47
Dialogic transformation of the moral 148–150
Dialogic digital ethics 174, 192, 196, 216
Dictatorship 8, 56, 97
Digital
  age 75, 176
  colonization 4, 10
  connectivity 126, 127

democracy x, 10, 11, 95
economy 126
emigrants 176
era 18
executive 29
hyperconnectivity 29, 31, 35, 53, 55, 75, 88, 89, 91–93, 95, 154, 156
innovation 170, 194
intelligence 203
paradox 42
phrenology 73
politicians 29
reality 6, 194, 196–198
representations 62, 67, 68, 70
researcher 29
resources 172
revolution 10, 168, 186, 195, 229
techniques 32, 33, 62, 63, 90, 117
technologies 5, 7, 10, 13, 18, 93, 117, 118, 126, 174, 176, 184, 186, 187, 193, 194, 220, 224, 237–239
transformation 29, 42, 43, 62, 72, 92, 95, 115, 117, 119, 126, 127, 150, 152, 156, 162, 192, 239
twins 62–71, 74, 75
utopia 171
Digitally hyperconnected citizenry 93, 152
Digitally hyperconnected society 33, 52, 98, 115, 139, 150, 154, 159, 161, 162
Digitization 10, 11, 33, 42, 53, 74, 89, 149, 171, 172, 194
Dignity x, 5, 9, 16, 20, 36, 56, 118, 160, 169, 193, 195, 196, 199, 201, 210, 212, 228
Dilemmas 6, 26, 104, 120, 153, 158, 186
Diminished cognitive 137
Disaffection
social 5, 84, 178
social responsibility 178
Discrimination 104, 172, 177, 178, 240
Discursive approach 149
Disinformation 96, 97, 171, 184–187
Disinterest 53, 116, 135, 137, 160, 186
Distrust 5, 6, 13, 42, 96, 116, 168, 201, 233
Diversity 19, 172, 196, 213

Division of powers 6, 8, 187
Donati, P. 33, 98, 117, 126
Doxa 83
Drive autonomously 34

E

Economic
  agents 126, 127, 140
  efficiency 6
  laws 6
  rationality 6
  rights 5
  spheres 33, 106, 109, 127, 133, 211, 212
  sustainability 6, 31, 62, 74, 118, 132, 236
Economy
  cryptoeconomics 127–133
  digital hyper-economy 126–140
Effectiveness 5, 19, 26, 31, 72, 94, 151, 175, 178, 183, 195, 197, 209, 211, 220–223, 231, 236
Election of representatives 11, 172, 180
Electoral fraud 64, 65, 96
Elitist democracy x
Emotional brain 180
Emotional harm 137
Emotions 6, 8, 45, 109, 114, 151, 153, 178, 180–185, 187, 217
Emotivist
  bias 11, 31, 52, 104, 160
Employment 5, 186, 230
Empowerment of citizens 179
Environmentally 131, 132
Environmental sustainability 19
Epistocracy 8, 187
Equality x, 5–9, 56, 83, 84, 149, 168, 213, 221, 222, 234, 239, 243
Equal participation 14, 16, 98, 159, 179, 196, 198, 202, 208, 214, 218, 221, 222, 235, 242
Equanimity 175
Ethereum 128, 129, 131, 132
Ethical
  algorithm 114

challenges 129, 139, 140
conflicts 105, 154, 242
dimension 76, 194
foundations of democracy 7
framework 105, 228, 238
governance of machines 117–120
patches 108, 111, 114
perspective 14, 18, 159, 160, 216, 220, 224, 235, 241
Ethical-critical perspective 53
Ethical-discursive perspectives 135
Ethic of discourse 150
Ethics
of democracy ix, 18, 228, 229
of discourse 150
Ethics Guidelines for Trustworthy AI 119
of institutional design 215, 216
line 74, 203, 237–243
lines and blockchain 239–240
lines and data mining 240–241
of robotic intelligence 108
Ethification 52, 152, 153, 157, 161
Evolutionary model 28
Exclusion 53, 54, 72, 84, 107, 172, 211, 212, 218
Expertocracy x, 6, 7, 178, 187
Explicability 74, 118, 161, 196, 198, 199, 223, 238, 239
Explicability Report 138, 162, 239, 242, 243
Extended
democratic 43
ethics 105
Extensive 49, 135

F

Fairness 55, 132, 138, 175, 196, 220
Field of distortion 96, 136
Financial crisis 5, 6
First era of AI 25–29
Formation
of the common 6
Foundation of morality 148

Framing algorithmic democracy 140
Franchise 61
Freedom 8–11, 56, 82–84, 96, 104, 118, 148, 154, 157, 161, 168, 172, 184, 193, 197, 208, 210, 217, 220, 221, 229, 234, 243
Functional dimension 219

G

Gender equality 172
General interests 13, 19, 82, 83, 211, 217, 244
Generating popularity 94
Generative AI 47, 50, 51, 67–70, 96–98, 154, 162
Generative Pre-trained Transformer (ChatGPT) 44, 50, 51, 96, 98, 153, 154
Generative adversarial network (GAN) 68, 69, 98
Gödel, K. 28
Good governance 75, 203, 238
Good practice 75, 118–120, 162, 235, 240
Google's *Advanced Technology External Advisory Council* (ATEAC) 119, 232
Google's *Neuronal Machine Translation* (NMT) 110
Governments 4, 8, 9, 13, 35, 42, 44, 47, 49–52, 56, 57, 66, 68, 72, 73, 75, 90–92, 94–98, 105, 106, 118, 150, 155, 171–174, 181, 182, 201, 203, 208, 215, 230, 231
Gramsci, A. 178
Guiding principles 218

H

Habermas, J. 12, 14–16, 19, 34, 35, 82–84, 89, 92, 98, 139, 149, 155, 156, 158, 159, 162, 163, 169, 178, 179, 183, 187, 196, 208–212, 221, 222
Health crisis 6
Hebb, D. 27
Hermeneutic
  perspective 213
Heteronomous society 115
Heteronomy 72, 90, 135, 137, 160, 171
Hidalgo, C.A. 63–68, 71, 73, 75
Homo data 127
Homo politicus 180
Horizon
  of action 7, 11, 16, 201
  of democracy 19

Human
activity 31, 33, 42, 62, 63, 90, 91, 148, 149, 151, 152, 173
behaviour 10, 108, 110, 128, 136, 152, 176, 194
being 10, 11, 27, 29, 35, 36, 43, 44, 46, 48, 51, 54–56, 96, 104, 112, 114, 116, 126, 127, 148, 160, 161, 172, 178, 181, 197
control 197, 241
dignity 36, 172, 194, 195
interlocutors 93
obsolescence 36, 57, 156
oversight 197
participation 197
reflexivity 176
relations 127, 148, 156
rights 36, 172, 195
Husserl, E. 168, 169
Hyper
hyper-AI 35
hyper-algorithmization 32
hyperconnected citizenry 93, 152
hyperconnectivity 10, 12, 30, 32, 53, 65, 69, 70, 72, 88, 126, 158
hyper-connectivization of the public sphere 44
hyper-datafication 32
hyper-datafied 156
hyperdigital society 149, 242
hyperdigitized economy 139, 140
hyperdigitized world 126, 162
hypereconomics 127
hyper-economy 11, 127, 129
hyperethics 52, 155–163
hyperethificiation 153
hyperhistory 127, 157
hyper-information 157
hypertextuality 157

I

Idea of justice 169, 224, 235
Ideology 41, 43, 52, 56, 96, 97, 154, 155, 169–178, 187, 193, 217
Ideology of dataism 176
Illiberal democracy 8, 97, 186

Immersive 133–137
Immoral patterns 137
Inclusion 7, 35, 67, 69, 74, 84, 120, 149, 150, 172, 196, 199, 213, 217, 238, 240, 242, 243
Independence 11, 175
Individualism 12, 152, 168, 243
Individual rights 6, 8
Inequality 5–7, 17, 42, 54, 84, 90, 96, 107, 131, 156, 160, 172, 177, 178, 184, 186, 187, 229
Influence public opinion 95, 185, 194
Infodemic 185
Information 10, 19, 31, 34, 36, 43, 45, 47–49, 52, 54, 55, 63, 66–68, 70, 74, 83, 88, 90, 92–94, 96, 106, 109, 112, 115, 127, 140, 149, 151, 155, 158, 161, 162, 171–173, 177, 183, 185, 186, 196, 198, 202, 208, 210, 228, 234, 236, 237, 239–242
Informed consent 154, 196
Infosphere 194, 196
Inner world 35, 133–135, 137, 140, 180
Institutional
  actors 215, 235
  design 10, 19, 96, 194, 200–204, 216, 218–224, 228–230, 233–235, 238, 243
  reification 229
  institution's legitimacy 220, 221
  institutions of civil society 4, 203, 210, 215, 216, 222, 234
Instrumentalization 36, 90, 113, 156, 160, 161, 171, 175, 223, 236
Instrumentalization of ethics 231, 232
Integrity 43, 53, 75, 175, 240
Intelligent mobility 104
Intelligibility 68, 161, 198, 199, 242
International guidelines 75, 198, 232, 236
Internet of Things (IoT) 30, 31, 43, 62, 168, 170, 193, 198
Intersubjective consensus 139, 149
Intersubjectivity 114, 139, 149
Intimity 34, 35, 110, 127, 135, 137, 140, 230
ION 44, 49
Irresponsibility 57

J

Jakarta PoW 132

Jevons, W.S. 128
Justice 5, 7, 10–12, 14–16, 35, 53, 82, 89, 93, 97, 104, 109, 118, 138, 149, 153, 155, 159, 160, 169, 170, 173–175, 177, 178, 183, 186, 187, 196, 209, 212–214, 216, 218, 220, 224, 229, 242, 244
Justify ix, 12, 13, 49, 56, 57, 107, 116, 150, 155, 179, 184, 185, 187, 194, 208, 219, 221, 222

K

Kant, I. 15, 82, 150, 168, 184, 200, 201
Keane, J. 29, 30, 34, 82, 96, 215, 217, 218, 238
Knowledge x, 9, 15, 29–32, 47, 66, 71, 82, 83, 89–91, 105, 106, 108–110, 116, 126, 130, 148, 149, 154, 158, 163, 168–170, 174, 176, 178, 183, 187, 192, 194, 198, 200, 202, 203, 208, 213, 218, 223, 240

L

Large Language Models (LLM) 49, 50, 67–70, 152, 154
Large-scale databases 89
Law 15, 30, 46, 48, 64, 65, 68, 82, 84, 151, 157, 179, 208–211, 222, 232, 233, 236
Leader Lars 44, 47–49, 106
Learning 27, 29–31, 49, 70, 71, 82, 104, 108–111, 113, 114, 116, 154, 173, 174, 194, 198, 239, 241
Learning by association 111, 113, 114
Learning through evolutionary selection 111
Lebenswelt 17, 98, 162, 212
Legal institutionalization 15, 208
Liberal
  democracy 4, 168
  foundations of democracy 8
Liberalism 6
Lifeworld 17, 19, 34, 89, 98, 139, 155, 156, 158, 163, 177, 217
Linkable things 127
Local spheres 5
Lovelace, A. 26, 27

M

Machine learning (ML) 27, 30, 44, 46, 57, 94, 110, 114, 115, 154, 194, 239, 241
Macro-deliberative approaches 19
Manipulation 8, 36, 72, 84, 93, 94, 96, 97, 171, 182, 185, 186, 223
Mass surveillance 73, 151

Masters of mankind 89
Mathematical models 10, 25, 28, 30, 31, 35, 45, 50, 52–57, 64, 72, 95, 98, 104, 106, 107, 109, 110, 115, 117, 126, 132, 150, 152–154, 157, 161
McCarthy, J. 27, 103, 104
Menger, C. 128
Meso-deliberative 18, 214
Meta Corporation 138
Metanomic agents 137
Metanomic ethics 135–140
Meta-user's experience 136
Metaversal activity 138
Metaversal platform 138
Metaverse 64, 132–140, 170, 183
Metaverse environments 137, 139
Metaversian algorithms 137
Metaversian avatar 134, 136
Metaversian economic 135
Metaversian frameworks 135
Metaversian systems 135, 138
Michihito Matsuda 46, 55
Micro-deliberative approaches 19
Model of democracy 12, 17
Modern democratic systems 43, 75
Moral
  agents 12, 150
  autonomy ix, 14
  awareness 3
  Choice Machine 153
  decisions 109, 116
  demand ix, 199, 213, 216
  disorientation 137
  justification x, 162, 201
  knowledge ix, 14, 16, 19, 109, 115, 117, 140, 148–150, 157–161, 163, 194, 197, 203, 208–211, 213, 215, 216, 221, 228, 231, 233, 243
  learning by algorithms 103–120
  machine 104, 105, 115–117, 153, 158
  problems 135, 138, 150, 194
  resources 18, 196, 197, 203, 209, 212, 216, 218, 232
  sentiments 137

sphere 12, 20, 52, 150, 152, 157, 162, 209, 212
strategy 199, 213
of the thermostat 103–108
validity 13, 18, 20, 73, 129, 139, 149, 157, 161, 195, 197, 199, 200, 208, 209, 212, 215, 229, 233
Morally intelligent technology 103–120
Morality of the thermostat 103–108
Moral machine learning (MML) 104, 105, 109, 111–116
Motivational scaffolding 111, 113

N

Neoliberal austerity policies 6
Neorealism 7
Neumann, J. von 27
Neuro
neuroculture 183
neurodata 178–184
neurodisciplines 180
neuroeconomics 110
neuroethics 110
neuromarketing 110
neuropolitics 110, 178–185
neuropower 178–184
neuroscience 30, 108–110, 180, 182, 183, 193
Neutral x, 6, 10, 51, 151, 153, 168–170, 173–177, 220
Neutrality 10, 12, 13, 97, 153, 155, 169–171, 173–175, 177, 187, 192–194, 229, 230
New
democracy 10, 52
economies 6
socio-political 32, 117, 161
NFT revolution 128–133, 135
Non-discrimination 196
Non-fungible tokens 64
Non-human interlocutors 93
Non-maleficence 118, 195
Non-political spheres 213
Normative dimension 218, 219

O

Objectivism 168, 169, 171, 192, 193
Objectivity 10, 97, 107, 153, 173, 175, 182
Opacity 54, 72, 75, 107, 198, 200, 203, 230, 232, 238
OpenAI 50, 96
Opinions x, 8, 11, 16–19, 36, 41, 42, 44, 45, 47–49, 51–54, 56, 64–66, 69, 70, 72–75, 82–84, 90–98, 106, 109, 114, 129, 136, 152, 153, 157–159, 162, 171, 183–185, 193, 199, 209, 211, 212, 214–216, 222, 224, 229, 231, 233, 237, 238, 244
Optimistic techno-determinism 156
ORACLE 154
Ortega y Gasset, J. 181
Outliers 75, 240, 241

P

Pandemic 5, 33, 67, 92, 151, 168, 185
Paradox of digitization, the 41–44
Partiality 56, 220
Participation ix, x, 4–9, 12–16, 18–20, 43, 44, 49, 53, 57, 63–66, 69–71, 74, 75, 82–84, 90, 95, 98, 120, 136, 158, 160, 162, 171, 172, 179, 193, 196–203, 208–210, 212–218, 221–224, 228, 233–240, 242–244
Participatory democracy 14, 19, 215, 216
Performance 9, 137, 169, 183, 197, 229
Personalized digital twins 66
Personal subjectivity 170, 173
Perversion 160, 161
Platform 13, 32, 44, 48, 50, 67, 70, 88, 89, 97, 104, 128, 153, 171, 209, 222, 229, 242, 244
Platformization 82–98
Pluralist democracy 6
Polarized digital 42
Policymaking 53, 56, 65, 66, 69, 91, 105, 106
Political 4, 30, 42, 63, 82, 106, 153, 169, 192, 208, 230
  autonomy x, 4, 7, 84, 179, 182, 193
  challenges 34–36
  clientelism 5
  colonization 18, 244
  corruption 46, 64, 65, 153
  decision-making 10, 35, 43, 49, 51, 52, 63, 66, 69, 91, 93, 192

decisions 11, 46, 52, 65, 67, 75, 83, 98, 177, 180
governance 7, 10
institutions 4, 42, 55
life 4, 181, 210
participation 5, 8, 16, 19, 62–64, 68, 70, 83, 180, 234, 237
parties 4, 5, 7–9, 18, 42–44, 48, 52, 67, 73, 91, 93, 95–97, 118, 169, 178, 179, 215, 216
power 4, 82, 210, 215, 234
processes 14, 15, 44, 66, 74, 159, 160, 179, 208
proposals 10, 35, 46, 53
responsibility 11, 17
rights 5, 8, 83
sphere 19, 43, 57, 72, 95, 106, 162, 209–211, 213, 215
system 5, 16, 42, 45, 65, 68, 71
Politicians 4, 5, 9, 33, 42–45, 48, 49, 51, 55–57, 65, 66, 92, 94, 95, 153, 168, 178, 187
Politics ix, 4, 9, 11, 15, 18, 19, 31, 32, 42, 44, 45, 47, 55, 57, 63, 65, 66, 71, 88, 91, 95, 96, 98, 118, 138, 152, 155, 168, 172, 178, 180, 181, 183, 187, 200, 215, 229, 237, 243
Populism 6–8, 84, 178, 181, 182, 184, 186, 187
Populist assembly democracy 6
Positivism 169–171, 174, 193
Post-democracy 4, 8
Post-truth 93, 172, 178, 184–187
Poverty 5, 72, 158, 163
Power x, 4–6, 8, 9, 11, 12, 14, 16–19, 28–30, 32–34, 36, 42–44, 48, 53, 68, 71, 74, 82, 89, 93, 95, 96, 98, 117, 126, 131, 136, 155, 156, 159, 160, 163, 168, 170, 171, 173, 175, 177–180, 182, 184–187, 194, 210–219, 221, 222, 224, 229–232, 243
Practical reason 161, 184, 209
Predictability 32, 68–70, 90, 110
Preferences 14, 33, 49, 51–54, 56, 64–67, 69, 70, 74, 75, 90, 105, 109, 126, 137, 138, 152, 153, 160, 174, 178, 180, 220, 243
Principle
of autonomy 154, 161, 196
of equality 238
of explicability 69, 184, 193, 197–204, 216, 223, 224, 241
of inclusivity 222, 238
of justice 154

of non-maleficence 154
of participation 198, 238
of publicity 82, 200, 201, 216, 223
of reciprocity 222
of universalization 149, 208
Privacy 35, 42, 114, 118, 120, 196, 230–232
Private interests 9, 218
Procedural moral principle 116, 149
Professional ethics 75, 119, 194
Proof of Hash Ownership 132
Proof of History 132
Proof-of-stake (PoS) 132
Proof-of-work (PoW) 131, 132
Prosocial emotions 116, 137
Prudent 139, 149, 241
Psychopathologies 90, 135, 136, 156
Psychosocial pathologies 72, 135, 136, 139
Public
commitment 50, 75, 202, 230–232, 234–236, 244
engagement 230–234, 236
laws 71
office 5
opinion 72, 73
policies 9, 11, 12, 31, 42, 44, 49, 54, 55, 90, 107, 151, 170–172, 177, 179, 182, 185, 192, 194, 242
response 186
responsibility 11, 17, 212, 215
spending policies 6
sphere 11, 44, 53, 82–84, 88–93, 95–97, 152, 162, 203, 208, 216, 221, 235, 238, 243
Public opinion 11, 15–17, 72, 73, 82–84, 91, 94–98, 171, 178, 183, 185, 208, 211, 212, 216
artificial 85, 91, 96
Published opinion 16

R

Rationalism 169, 181
Rationally criticize 138, 150
Reciprocal

expectations 149
recognition 12, 13, 186, 201, 210, 220, 228
Reconstructive methodology 208, 213
Rectitude 175
Reductionism x, 11, 176, 183
Regional spheres 5
Regression ix, 4, 156, 160, 161
Regulation 19, 20, 42, 65, 67, 70, 118–120, 182, 195, 201, 209, 211, 222, 228, 230–234
Reification 34, 35, 90, 160
Reinforcement learning 70, 111, 112, 114
Reliability 69, 120, 195, 239
Representative
democracies 7, 15–17, 52, 64–66, 215
system 5, 11, 13, 16, 42, 172, 216
Representativeness 10, 69, 70, 232
Responsibility 10, 12, 14, 36, 43, 48, 57, 72, 74, 112, 115, 118, 127, 132, 155, 157, 160, 162, 172, 177, 178, 181, 196, 201, 210, 211, 215–217, 228–230, 232–234, 238, 239, 243
responsibility research and innovation (RRI) 120
Revisionism 7–9, 12, 13, 15, 187
Roboethics 108
Robotic 34, 108, 118
Robust 91, 105, 132, 196, 241

S

Scalability 69, 70
Scientism 171, 193
Scrutability 69, 70, 74
Searle, J. 27, 28, 103, 104, 117
Second Age of Artificial Intelligence 29–32
Self-beliefs 103
Selfishness 11, 46, 52, 54, 109
Self-regulatory 119, 120, 231, 236
Semantic analysis machine (SAM) 42, 44–46, 51, 55, 106, 153
Sense of justice 36, 177, 184
Simon, H. 27
Singularity 117, 153
Social

bots 72, 73, 92–98
contract 5
credit programme 73
development 5, 32, 42, 75, 117
entropy 74
injustice 6, 154, 158, 162, 163
justice 5
life 117, 127, 171, 176, 209, 213
media 9, 45, 46, 49, 51, 56, 67, 73, 92, 95, 96, 176, 177, 187, 231
pathologies 184
politician 94
responsibility 108, 192, 203, 209, 230, 236, 237
responsibility reports 203, 235, 236, 242
rights 5
services 5
Socio-political sphere 33
Spaces
of participation 93, 215, 235, 244
of relationality 14, 139, 149
of trust 214, 235, 237
Specific codes 114
Spheres
of activity 32, 148, 149
of civil society 16, 18, 20, 209, 210, 212–214, 228
of democracy 105, 116
of participation and communication 238–242
Stakeholders 19, 54, 132, 196, 197, 199, 202, 203, 232–234, 236–238, 240
State 6, 35, 45, 68, 84, 103, 151, 179, 196, 208, 228
Statistical models 175
Stephenson, N. 133
Strategic
action 18, 20, 155, 212, 214, 217
rationality 8
Strong AI 27–29, 113
Substitution of the citizenship 12
Surveillance capitalism 126, 132
Sustainability 6, 31, 32, 62, 74, 90, 118, 132, 232
Sustainability reports 203, 235, 236
Sustainable Development Goals 230, 236

Sustainable economic growth 192
Symbolic model 27
Symbolic production 98, 156
Synchronicity 69, 71
Synthesising moral justification 162
Synthesising reality 162
Synthetic
  datasets 97, 162
  ethics 162
  opinion 97, 162
  party 47, 48
  reality 98
  sphere 81–98, 219
Synthetification 162
Systems of representation 179

T

Tay 55–57
Technical ideology 171
Technological
  colonization 92
  gap 53
  rationality 192
  revolution 177
  singularity 27, 29, 117
  solutionism 171
  sphere 42
Technology x, 10–12, 30, 32–34, 42, 45, 47, 48, 50, 53, 55, 56, 62–65, 67–72, 89–91, 96–98, 117, 118, 126, 134, 135, 137, 140, 151, 156, 158, 160, 168–171, 173, 174, 176, 177, 182, 187, 193–195, 197, 198, 202, 203, 211, 217, 222, 229, 231, 232, 239, 242, 243
  companies 50, 55, 75, 97, 193, 210, 229, 230, 234, 243
  corporations 132, 203, 229, 232
  platforms 98, 209
Technopopulism 7, 9, 12, 184–187, 215, 243
Technocratic expertocracy 187
The Guardian 92
The Los Angeles Times 106, 107
The New York Times 130

Tokenization
of the economy 138
Trust 4, 5, 8, 13, 20, 31, 33, 35, 51, 55, 82, 93, 116, 118, 120, 132, 155, 173, 174, 177, 178, 180, 181, 183, 187, 192–198, 200–204, 211, 213, 216, 218, 221, 223, 228–239, 241–244
Trustworthy 18, 83, 193, 195, 196, 202, 224
Turing, A. 26, 27, 103
Two-Round PoW 132
Two-track democracy 13–20

U

Universalist ethics x
Utilitarian approach 10
Utilitarianism 149
Utility functions 112, 113, 116

V

Validation Tool for Advanced Life Sciences (VITAL) 106
Valid interlocutors 16, 184, 195, 209, 210
Value accretion 112
Value learning 113
Values x, 4, 5, 7–9, 12–14, 17–19, 35, 36, 55, 64, 72, 74, 83, 89, 104, 109, 111–117, 120, 127–130, 132, 134, 136, 139, 140, 148–150, 155, 156, 158, 160, 162, 163, 170, 173–176, 179, 181, 183, 185, 194–200, 203, 209, 210, 215, 218, 219, 222, 228, 229, 231, 233, 235–238, 241, 242
Violence (Gewalt) 55, 56, 73, 89, 129, 156, 217, 221, 223
Virtual environment 138
Virtualized other 127
Virtualized public sphere 157
Virtual politician 44–52, 55, 64
Virtual reality (VR) 132, 133, 136–138, 140, 170
Virtual space 134, 135, 137
Virtual world 70, 127, 133, 139, 140
Virtues 36, 74, 116, 139, 148–150, 155, 157, 158, 162, 163, 176, 228
Vulnerability 90, 94, 115, 118, 126, 135, 139, 148, 150, 162, 228

W

Walras, L. 128
Weak AI 27, 114, 117
Web 2.0 88

Whistleblowers 75, 239, 240

X

Xenophobic 55, 56, 72, 111, 119, 137, 160
Xian’er 152

Y

Yāffar, A. 25, 26
Yandex 55, 153

Z

Zamagni, S. 29, 33, 34
Zuckerberg, M. 132, 231