Introduction

Thinking Outside and Inside the Box to Govern AI

Artificial intelligence (AI) will cure cancer. AI will destroy humanity. AI is the *ichor* of humanoid robots. AI is Big Data. AI is the new oil, the new electricity, the new Industrial Revolution, the new atomic bomb. AI is like tobacco. AI is like junk food. AI is a parrot.[1] AI is a mirror.[2] AI is marketing hype – snake oil.[3]

I am sure you have read or heard these statements about AI.[4] Depending on which opinion leader you follow, you might think that it is the most transformative technology of our times, a hyped-up technology that has not given humanity reliably useful products yet, or something in between. In popular imagination, AI is at once magical and dangerous. Crowd-pleasing movies portray it as a soulless destroyer (*The Terminator*), a soulless seducer (*Her*), or a misunderstood, curious humanoid on the verge of developing consciousness – and, who knows, a soul (*Atlas*). Novels like Kazuo Ishiguro's *Klara and the Sun* and Ian McEwan's *Machines Like Me* interrogate what it means to be human when machine intelligence is so close to our minds, hearts, and perhaps supermarket shelves. Documentaries like *Coded Bias* and *The Social Dilemma* take issue with the social ills associated with software and platforms powered by algorithms. It seems like we are at the precipice of something great if we get AI right. And if not … well.

This book is about efforts to get AI right. Building upon AI scholarship in science and technology studies, technology law, business ethics, and computer science, it documents potential risks and actual harms associated with AI, lists proposed solutions to AI-related problems around the world, and assesses their impact. This book presents as many theoretical debates and as much empirical evidence as possible to

[1] Bender et al., "On the Dangers of Stochastic Parrots."

[2] Vallor, *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*.

[3] Narayanan and Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*.

[4] A note on term usage: Throughout this book, *AI* refers to the underlying science and technology. Qualifiers, such as *generative* or *symbolic*, are used where appropriate. *AI tools* are software, services, or platforms available to customers. *AI systems* denote the combination of hardware and software systems with AI and non-AI components. Specific definitions are provided in Chapter 1.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001

document how and how well *technical solutions*, *business self-regulation*, and *legal regulation*, which stand out as AI governance toolkits, work.

Governance refers to coordinating and steering a society with a multiactor, multilevel, and networked framework.[5] Whereas standard theories of government focus on the state as the organized actor ruling over society, governance models argue that international organizations, nation-states, local governments, businesses, and nongovernmental organizations coordinate and decide in harmony. Thus, governance acknowledges multiple levels of jurisdiction and a networked decision-making process rather than top-down decision-making by government.[6] *AI governance* is about collective decisions about people, and the artifacts they develop and use with an AI component.[7] Such decisions come from businesses developing and using AI systems, regional, national, and subnational governments and, to a lesser extent, international bodies. As it will become obvious, AI governance has not emerged as a result of conscious coordination by multiple actors; to the contrary, a combination of technical fixes, business practices, and government regulation has coalesced in an unplanned and imperfect way. Throughout this book, *AI governance* is used as a descriptive, not normative, term. In other words, governance does not a priori mean that the technical, business, and government solutions to AI risks and harms are always useful and effective. Rather, the objective in this book is to hold this governance model to critical scrutiny.

AI can be governed partially through a combination of scientific and technological improvements in accuracy, reliability, safety, data protection, and fairness; in-house or external boards, councils, and teams that oversee adherence to ethical business principles in the AI sector; laws and policies regulating AI products (e.g., chatbots and automated decision-making software) and techniques (e.g., deep learning and generative AI) specifically; and laws and policies addressing issues adjacent, but not exclusive, to AI, such as data privacy, data protection, antitrust, and content moderation. In the ideal scenario, technical solutions propel AI models and resulting products to successful implementation of ethical norms, while self-regulatory bodies commit businesses to ethical conduct, and laws and policies set binding parameters around private-sector practices to ensure respect for fundamental rights. This idealized governance scheme is likely to produce some desirable results in part because AI-centric and AI-relevant laws and policies build upon established legal systems with a history of regulating the negative impact of scientific and technological developments, and in part because business incentives are at times aligned with the deployment of accurate, reliable, safe, and fair AI products.

[5] Bevir, *The SAGE Handbook of Governance*; Rhodes, "The New Governance: Governing without Government."

[6] Kooiman, "Social-Political Governance: Overview, Reflections and Design."

[7] For recent scholarship on AI governance, see: Bullock et al., *The Oxford Handbook of AI Governance (Oxford Handbooks)*; Batool, Zowghi, and Bano, "AI Governance"; Veale, Matus, and Gorwa, "AI and Global Governance: Modalities, Rationales, Tensions."

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001

The idealized governance model will not work seamlessly, however. Each component of this governance scheme has weaknesses – some more than others. Technical solutions are often found in proof-of-concept papers, with little evidence of business adoption. Even when they are implemented, there is no evidence that they offset the magnitude of the problems they claim to address. Ethical AI boards, councils and teams, at arm's length from the core money-making endeavors they oversee, have been fighting for relevance and impact, if not survival, since their origins in the mid-2010s. In other words, technical solutions and business self-regulation have limited leverage vis-à-vis the business incentives, models, and practices they want to reform. Disciplining businesses through binding laws and policies remains an untapped potential: There are numerous bills but few actual laws regulating AI. Legal regulation is not without its shortcomings, though. Lawmakers in major AI-producing countries are all too happy to seek advice from the same corporations they are supposed to regulate; thus, the risk of what legal scholars call "regulatory capture" by industry is all too real. The systemic negative impact of AI on the environment, workers' rights, and market competition has received almost no legal attention. Furthermore, some of the worst risks of AI – for example, lethal autonomous weapons systems – keep concerned activists, journalists, academics, and international lawyers up at night, but politicians have averted serious regulatory proposals in national or international law.

The shortcomings of techno-solutionism, business self-regulation, and legal regulation are only part of the story. Getting AI governance right cannot be disentangled from local and global struggles for justice and equality. A racist society cannot but develop racist AI. A sexist society cannot but develop sexist AI. A society in which a significant part of the population denies basic facts and scientific findings cannot get content moderation right. A society in which technological progress empowers only a small number of corporations – and a few startups in their orbit – cannot build fair AI. Critical decisions around AI governance (e.g., decisions on what constitutes hate speech, who deserves to be hired, what to do when an algorithm flags the wrong person for arrest, or even what is true and what is false) are almost entirely made by the companies developing AI products or government and private-sector actors using them. The people who are negatively affected by those decisions are denied a seat at the table. It is true that AI creates new problems and amplifies existing ones, but deep down, where AI fails is where human societies fail.

Data-driven, machine learning-based AI was born into a planet facing an ecological crisis, declining trust in political institutions and severely challenged information ecosystems in democracies, not to mention deepening national and global inequality. Most people, including citizens of technologically advanced countries, feel alienated from the decisions shaping their lives. The gap between citizens and the politicians who are supposed to represent them is widening. At any rate, a small number of unelected corporate leaders have been making decisions that shape people's lives as much as, if not more than, governments. People enjoy the convenience

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001

and fun brought by contemporary technologies but cannot stop asking why supposedly transformative technologies come with a baggage of bias, discrimination, surveillance, disappearance of data privacy, disinformation, rumors of massive unemployment, and the specter of even more inhumane warfare than ever before.

That is why this book is a call to think inside *and* outside the box. Pragmatically, existing AI governance models are too important not to advocate for. Technical solutions, business self-regulation, and – especially – legal regulation can mitigate and even eliminate some of the potential risks and actual harms arising from the development and use of AI, and no other short-term solution is realistic. However, the long-term health of the relationship between technology and society depends on whether ordinary people are empowered to participate in making informed decisions to govern the future of technology – AI included. Put simply, the problem is disempowerment and alienation, and the solution is people's informed participation.

Reflecting on decades-long critical discussions on modern technology, I argue that the development of today's AI is at once a cause and consequence of most people's disempowerment and alienation. Academic and journalistic evidence of AI risks and harms refutes the naïve assumption that all technologies benefit everyone without qualifiers. Misalignment between the profit motive and human interests, the presentation of artificial intelligence as an expert field, the market concentration of a few companies, research and development gaps that threaten to deepen access inequality within and across countries and the hierarchies of the capitalist workplace all contribute to disempowerment and alienation. Politicians and political parties have so far done little to close the gap between the interests of most people and decisions made by AI industry leaders. Academia, suffering its own crisis of alienation from the public, faces stiff competition from industry in research infrastructure and recruitment of top talent. Nongovernmental organizations, academics, journalists and a few politicians advocating for the defense of rights in the context of technological change appear to be the only actors bringing ordinary people's struggles to public debates, but their economic and political power is limited.

Anyone interested in the development of AI as a technology in the service of the common good, in whatever way defined, should therefore take the call for the expansion of citizens' agency in making collective decisions seriously. The technical intricacies involved in technological decision-making cannot be ignored, but the academic and journalistic evidence compiled in this book leaves no doubt that decisions that go into the research, design, development, deployment, and use of AI are about the distribution of benefits and risks among people – in a word, politics. Thus, I argue that everyone has an interest in, and skills for, shaping the future of AI.

This analysis of technical, corporate, and legal governance should acknowledge the true heroes in this story: Journalists, researchers, social movement activists, whistleblowers, lawyers, conscientious corporate workers, and (a few) politicians who have raised awareness around AI risks and harms. Every single case of course correction described in this book was made possible thanks to someone who sounded the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001

alarm after meticulous research on an AI system that affected someone negatively. Without them, we would have lived in a world of illusions in which we would accept without evidence that every technological product is good and anyone raising objections is a hopeless technophobe.

I believe AI can still be reimagined and developed as a technology that serves humanity while causing minimal or no harm. Humans can live in harmony with technology. It requires a robust governance model that leverages technical, business, and government solutions to detect and mitigate risks and harms. A genuinely successful governance model that goes beyond avoiding the worst harms to actually reconstruct AI as a public good necessitates a radical rethink of technology, and through it, society. Citizens cannot allow themselves to be passive recipients of technology in the age of AI. We need to reclaim our agency to build a better world.

DATA AND METHODS

This book draws upon a study of online media archives, which cover (1) newspapers (e.g., *The New York Times*, *The Washington Post*, *The Boston Globe*, *The Globe and Mail*, *The Guardian*, *The Economic Times*), (2) technology magazines (e.g., *Wired*, *TechCrunch*, *MIT Technology Review*); (3) online technology blogs; (4) academic publications in the social sciences and humanities, law, journalism, and computer science; and (5) company blogs and reports.

News articles about the following are selected:

- AI and AI-related concepts, such as machine learning, deep learning, algorithms, automated decision-making systems, and generative AI;
- Potential risks and concrete harms arising from AI, such as bias and discrimination, disinformation, surveillance, data privacy violations, and labor exploitation;
- Terms associated with socially desirable AI, such as accuracy, ethics, fairness, responsibility, safety, and transparency;
- Proposed technological solutions, such as AI for social good, de-biasing, AI detection, and red-teaming;
- Business oversight boards, councils, and teams, such as Meta's Oversight Board and Microsoft's Responsible AI Council;
- Bills and laws that aim to regulate AI or AI-related concerns, such as privacy and consumer protection, around the world; and
- Opinion pieces that offer a bird's-eye view of the relationship between AI and society.

The collected information is analyzed using quantitative and qualitative content analysis. The quantitative analysis relies on a comparison of term counts over time (e.g., the use of *AI ethics* over time), term co-occurrences (e.g., *de-biasing* used in the same sentence as *discrimination*) and in comparison to similar terms (e.g., *AI*

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001

ethics vs. *responsible AI*). These comparisons present a picture of the evolution of ideas and practices around AI governance. The qualitative component of analysis identifies common patterns, themes, and meanings, and instances of scholarly agreement and disagreement across texts. Every piece of academic and journalistic evidence that a proposed solution has worked or failed is documented, and analyzed in light of available information. If the explanation for an event or decision has generated disagreement among academics, journalists, social movements and industry spokespersons, multiple evidence-based viewpoints are reported.

BOOK OUTLINE

Chapter 1 introduces basic terminology. Terms such as *artificial intelligence*; *data*; *algorithm*; *machine learning*; *neural networks*; *deep learning*; *large language models*; *generative AI* and *symbolic AI* are presented to develop a sense of what AI is, how it has evolved and what it does. This chapter also introduces some of the major conceptual disagreements in the field. Different ideas about how to develop AI in the best way drive disagreements, as well as philosophical differences over what intelligence means and whether machines can develop human-like intelligence.

Chapter 2 is devoted to AI ethics, broadly defined. It provides an overview of *ethical*; *responsible*; *safe*; *trustworthy*; *transparent* and *explainable*; *accurate*; *just* and *fair*; *accountable*; *sustainable*; *robust*; *accessible* and *inclusive* AI. Just as the definition of AI itself is fraught with disagreement, words with a connotation of "good" AI have generated considerable controversy among academics, social movement activists, journalists, business leaders, and lawmakers. This chapter aims to represent the plurality of positions. Furthermore, the adjectives associated with getting AI right are mutually supportive, but tensions between desirable goals are mentioned as well.

Chapter 3 presents the other side of the coin, namely AI risks and harms. Automated decision systems, chatbots, recommender systems and other AI-powered software and platforms have been found to cause potential risks or actual harms to affected persons and communities. Such risks and harms include *bias and discrimination*; *surveillance*; *inaccurate, incorrect, and unreliable output*; *disinformation, misinformation, or manipulation*; *harm to life, livelihood and wellbeing*; *privacy violations*; *decline in product and service quality*; *political polarization, online radicalization and algorithmic censorship*; and *job replacement*. Some of these harms, like bias and discrimination, have already been experienced frequently, while others, like job replacement, point to future risks. It is also worth noting that AI risks and harms often aggravate existing social and political problems. For example, political polarization and radicalization, while exacerbated by algorithmic curation, appear to have origins in societal divisions. Finally, AI is criticized for causing system-level harm in the form of *environmental degradation*; *exploitation of labor*; and *market concentration*.

Chapter 4 is devoted to technical solutions to rectify AI risks and harms. *AI for social good projects*; *human-in-the-loop solutions*; *de-biasing*; *AI-generated text, image*

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001

and voice detection; and *testing* are presented as potential technical fixes. Detecting AI-generated content remains a major challenge. Human-in-the-loop solutions and testing have proven to be such commonsense practices that they are promoted by AI-producing or -using businesses themselves as well as by laws. AI for social good projects and de-biasing produce positive impact, but there seems to be a gap between expectations and reality. As is documented throughout this book, the roots of AI risks and harms are not technical; therefore, technical solutions cannot bring about transformative change in the face of AI risks and harms.

Chapter 5 addresses business self-regulation as an AI governance model. Voluntary AI principles and codes of conduct have risen to prominence in the absence of AI laws since the mid-2010s. Numerous large companies have established internal or external advisory boards or councils and responsible AI teams to hold themselves accountable. The evidence on these self-regulatory bodies is mixed: Journalistic reports suggest improvements in business conduct in a number of cases, but one cannot ignore the fact that none of the boards, councils, or teams can force businesses to respect their decisions or suggestions. What is worse, some powerful AI companies have ignored calls to create self-regulatory institutions or disbanded them at first sight of friction.

Chapter 6 is about laws as binding mechanisms to eliminate or mitigate AI risks and harms. Most countries have AI-promotion strategies that devote little or no attention to potential problems. The number of bills proposed in national legislatures to address those problems has been increasing since the late 2010s, but only the European Union and South Korea have thus far legislated laws regulating AI. Despite the absence of AI-centric lawmaking, however, some trends are emerging. First, AI regulation has been taking place, to a limited extent, in AI-adjacent realms such as data privacy and protection, consumer rights, antitrust, and children's protection. Second, the European Union's AI Act has set the trend for risk-based, future-proof, and technology-neutral legislation that will likely be followed by other countries. Third, the absence of national legislation in the United States, home to most cutting-edge AI technologies from the 1990s to the early 2020s, has led states and cities to take initiative. And finally, even the successful passage of a law does not address all AI risks and harms – lawmakers' omission of military AI as an area of regulation is a case in point.

Chapter 7 zooms out of conceptual and empirical studies of AI governance to ask if we can build a better future with AI. The technical, corporate, and legal governance models presented in this book are necessary but insufficient to endow ordinary people with the power to push back against risks and harms, and chart a course for AI for the common good. Thinking together with philosophers and social scientists in the Critical Theory, Science and Technology Studies, and Democratic Theory traditions, I argue that most people's experience with AI is one of fear as a result of their long-standing disempowerment and alienation from the technologies shaping their lives. Attributing disempowerment and alienation to technical aspects of AI

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001

is wrongheaded: It is the evolution of modern capitalism that has widened the gap between people and the technologies that are supposed to make their lives better. Reorienting the relationship between people and AI requires a radical-democratic politics that questions hierarchy in government and in the workplace. Technology can serve as a force for the social good only if informed citizens participate in the decisions shaping their lives in the design, development, deployment, and use of modern technology, AI included.

The prologue fleshes out the lessons drawn from this book. It offers best practices for a workable AI governance model that uses technical solutions, business self-regulation and legal regulation. Then, it delves into some of the shortcomings of that model. The radical-democratic perspective I advocate makes five general, practical suggestions for everyone concerned with AI risks and harms. (1) *Organize*: Build networks of support and civic organizations around technology-specific concerns as well as conventional rights considerations; (2) *Learn*: Acquire cross-disciplinary capabilities on the uses, practical applications, potential risks, and governance models associated with technologies like AI; (3) *Participate*: Push politicians and businesses to expand the boundaries of decision-making in the public and private sectors; (4) *Care*: Approach technological change from the perspective of vulnerable populations, and with an ethic of non-domination that refuses to treat nature and other people as instruments; and (5) *Resist*: Maintain an openness to contention with the producers and users of technologies that generate risks and harms.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 15 Nov 2026 at 04:31:39, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009738347.001