An Introduction to the Law, Ethics, and Policy of Artificial Intelligence

Nathalie A. Smuha

BEYOND THE AI HYPE

Artificial intelligence (AI) was founded as an academic discipline almost 70 years ago, when a conference took place at Dartmouth College. The proposal submitted by the conference conveners described the project as an attempt "*to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.*"[1] Just a few years before the Dartmouth Conference, Alan Turing had already published a paper titled "*Computing Machinery and Intelligence*," in which he kickstarted not only a philosophical discussion on whether machines could *imitate* human thinking but also discussed the development of digital computing and "learning machines."[2]

Over the years, significant advances toward the achievement of those aims were made. Periods of great optimism (so-called "AI springs"), during which the technology knew rapid advancements and attracted elevated levels of funding, were followed by periods of pessimism in the technology's progress (so-called "AI winters"), during which interest and investment in the technology plummeted, with a low point in the 1990s. Gradually, the wider availability of data, advanced computing power, and significant research progress (especially in the subfield of machine learning) contributed to AI's latest boom. Interestingly, "*from 2010 to 2021, the total number of AI publications more than doubled, growing from 200,000 in 2010 to almost 500,000 in 2021.*"[3]

[1] John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955.
[2] Alan Turing, "Computing machinery and intelligence" (1950) *Mind*, 59(236): 433–460.
[3] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault, "The AI Index 2023 Annual Report," *AI Index Steering Committee, Institute for Human-Centered AI, Stanford University*, April 30, 2023.

The current AI spring is explained not only by the increased uptake and normalization of AI applications across virtually all sectors of the economy but also by the advent of generative AI and other applications which found their way to the public at large, resulting in a true "AI hype." One can only speculate about whether this hype will soon (or has already) hit its peak and an AI winter is coming, or whether more breakthroughs are underway.[4] There are, however, many more important questions to formulate and points to make, which are not always raised in many of the brief summaries about AI's hype – points that may be overlooked precisely because of our enthusiasm for the perceived benefits of this impressive technology. Let me focus on three aspects in particular that deserve our attention.

A Long History

First, it should be born in mind that the history of AI as a *concept* dates back at least to antiquity, where myths already existed about "automata" or self-operating machines displaying human behavior.[5] Hephaestus, the Greek god of artisans and blacksmiths, was for instance said to have created an artificial man of bronze, Talos, to protect Europa – a Phoenician princess after whom the European continent was named – against potential invaders and kidnappers. Moving from myth to reality, Ancient Greece also saw the birth of the Antikythera mechanism – a hand-powered mechanical model of the solar system developed around 200 BC and used to predict astronomical positions, often described as the first example of an analog computer.[6]

The human drive to transgress the boundaries of the natural and the artificial and to create "intelligent" machines by no means diminished in the Middle Ages. For instance, in 1206, Ismail al-Jazari, an Arab polymath from Mesopotamia who is described as the "father of robotics," wrote the *Book of Knowledge of Ingenious Mechanical Devices*, including detailed accounts of how to construct musical robot hands and drink-serving waitresses.[7] Scientists started experimenting with the creation of mechanical devices for a range of purposes, sometimes even purposely inflating the machine's capabilities and misleading audiences (like the example of the Automaton Chess Player or *the Mechanical Turk*, which was actually controlled by a human operator sitting inside it).[8]

[4] The hype's bubble is also increasingly being pierced, as AI developers not always able to deliver the technology's promises. See also Eric Siegel, "The AI hype cycle is distracting companies" (2023) *Harvard Business Review*, June 2, https://hbr.org/2023/06/the-ai-hype-cycle-is-distracting-companies.

[5] See for example Silvio A. Bedini, "The role of automata in the history of technology" (1964) *Technology and Culture*, 5(1): 24–42.

[6] See also John Hugh Seiradakis and M. G. Edmunds, "Our current knowledge of the Antikythera mechanism" (2018) *Nature Astronomy*, 2: 35–42.

[7] See for example Shahino Mah Abdullah, "Intelligent robots and the question of their legal rights: an Islamic perspective" (2018) *ICR Journal*, 9(3): 394–397.

[8] See also Elizabeth Stephens, "The mechanical Turk: a short history of 'artificial artificial intelligence'" (2022) *Cultural Studies*, 37(1): 65–87.

In sum, humans have been fascinated with artificial beings long before the Dartmouth Conference, which is also evidenced by literary works, from the Golems of Chełm and Prague to Mary Shelley's Frankenstein's Monster. The question is then: how can we avoid that this historical fascination does not make us overly focused on what AI *could* do instead of reflecting on what it *is* doing and what it *should be* doing in practice? For it is precisely within the gap between *is* and *should* that many problems around the technology's development and use can be situated, including the nonchalant, negligent, or even malicious launch of problematic AI applications, from which harmful consequences can ensue.

One of Many Technologies

Second, it must be noted that AI is but one of many technologies, and numerous other innovations have preceded its hype and discourse. The history of technology counts a long list of inventions that were heralded as groundbreaking and that transformed our societies to greater and lesser extents. AI is being treated as a shiny new toy, and is sometimes even compared with the discovery of fire, electricity, oil, or nuclear technology, which has led experts to debate whether these analogies are useful (or to claim that none of them makes much sense). Yet the fact that such analogies are being made in the first place should serve as a reminder that "*what has been will be again, what has been done will be done again; there is nothing new under the sun*."[9] Human beings have always sought to deploy (new) tools in ways that serve their purposes, in good, bad, and negligent ways – and this certainly applies to AI too, as it is developed and used by human beings.

Society has dealt with many other (powerful) inventions in the past, and there is a rich history of (failed and successful) governance practices that can be dug into to analyze which lessons to draw when it comes to AI – and how to govern human behavior in relation to AI. While it may be tempting to treat AI as an entirely novel and different phenomenon stemming from human ingenuity, this attitude not only feeds an excessive hype but also risks overlooking the ingenuity that humans have shown throughout history when it comes to setting up mechanisms and institutions to govern society. It is, furthermore, a convenient position for those actors who would prefer *not* to draw any lessons from past governance experiences, as some seek to avoid AI-related governance measures altogether.

The hype-fueled fixation on AI as fundamentally distinct from other technologies also has two other problematic corollaries. In first instance, it reinforces the narrative that AI is something elusive and inevitable, manifesting itself in society in a form that we cannot quite grasp, and that cannot properly be defined or understood. But discussing AI as something abstract, ephemeral and almost magical overlooks its very concrete – and governable – building blocks, from software code and

[9] Ecclesiastes, 1:9.

data-filled Excel sheets, to physical CPUs, motherboards and data centers, and of course the human beings creating and operating them. In addition, it also overlooks the fact that other technologies which may not fall under the contours of AI can lead to equally impactful and problematic consequences, and that the focus should hence lay not (merely) on the technology but rather on the values society cherishes and wishes to protect. The question is, hence, how to avoid the trap of treating AI as entirely novel, while at the same time being sufficiently mindful of the very concrete ways in which it can (adversely) affect society, and ensuring tailored governance mechanisms to counter potential harms.

Societal Impact

Third, as already alluded to above, there is an important societal dimension that needs to be considered within any AI history, as it does not stand separate from its technological dimension. Like all technologies, AI is inherently embedded in society, thus affecting and being affected by the broader environment in which it is designed, developed, and deployed – for better and for worse. The societal impact of Artificial Intelligence is of course more noticeable the more it is being implemented and used in a diversity of domains, which also explains the relatively recent surge of (academic and other) interest in AI ethics, in parallel with the technology's increased uptake. Yet AI's uptake is also enabled and furthered by the societal condition. These enabling factors pertain, inter alia, to society's belief in innovation as an almost absolute good, its technology-solutionist orientation, and its conception of "progress" as almost coinciding with technological advancement rather than also considering if and how these advancements translate into higher individual and societal welfare – for all. Yet if we shape technology and technology also shapes us, it is essential to ask how it can be ensured that this mutual shaping process takes place in a way that protects rather than undermines our legal, moral, and political standards.

The attentive reader will have noted that the three questions I formulated above are all variations of the same theme – one that lies at the heart of this book: given that AI systems are increasingly being developed and deployed in ways that impact our lives, what role do *law*, *ethics*, and *policy* play to govern this impact and to ensure that the core values of society are safeguarded? Answering this question requires a cross-disciplinary lens, as it is only by looking at it from different perspectives that AI's societal effects can be grasped.

To this end, in the summer of 2021, I convened the first edition of the Summer School on the Law, Ethics and Policy of Artificial Intelligence at the KU Leuven Faculty of Law and Criminology. This program brought together a multidisciplinary group of lecturers and participants to the city of Leuven for an intense deep dive into a range of topics related to the impact of AI on society, with a particular focus on Europe.

Many of the chapters of this book were born out of the rich exchanges and discussions that took place within the margin of the first and subsequent editions of the Summer School. The purpose of this book is to consolidate those insights and make them available to a wider readership.

BOOK OUTLINE

This book addresses the main challenges and opportunities of AI not only from a horizontal perspective (covering general areas in which the advent of the technology raises questions, such as philosophy, ethics, and various legal domains) but also from a vertical perspective (considering AI's implications in a range of sectors), with the aim of providing the reader a more holistic understanding of AI's impact across society. Just like in the program of the AI Summer School, the primary jurisdiction discussed in the chapters concerns Europe, and the underlying societal model that is taken for granted is one that seeks to protect human rights, democracy, and the rule of law – three core values of constitutional liberal democracies.

The book's focus not only lays on the latest wave of AI applications but also encompasses discussions of more traditional algorithmic systems that are equally able to raise challenges to societal values, and that should not be overlooked merely because we have become so accustomed to them that they are now considered too "traditional" to be called "AI." Each chapter is self-standing, yet many of the themes discussed therein are recurring, in particular the acknowledgment that more interdisciplinary research and cooperation on AI is needed. The book is divided into three parts, each focusing on a different angle.

Part I: AI, Ethics and Philosophy

The first part of this book starts by conceptualizing AI as a scientific discipline and setting out its technical foundations. In Chapter 1, Wannes Meert, Tinne De Laet, and Luc De Raedt provide a perspective from the field to describe machine learning and machine reasoning, two domains within the broader field of AI that are rapidly evolving. They distinguish different types of functions and techniques, and close with some reflections on what it means to build "trustworthy," "explainable," and "robust" AI, thereby also building a bridge between their technical discussion and the book's subsequent chapters, which discuss AI from a philosophical lens, with a particular focus on moral philosophy or ethics.

Chapter 2, written by Vincent C. Müller, offers a structured overview of the philosophy of AI. After describing a broader set of AI definitions beyond computer science, he introduces the concepts of intelligence and computation, as well as the main topics of artificial cognition, including perception, action, meaning, rational choice, free will, consciousness, and normativity. Through a better understanding

of these topics, he argues, the philosophy of AI contributes to our understanding of the nature, prospects, and value of AI. At the same time, he also explains that these topics can be better understood by discussing AI, and thus suggests that "AI Philosophy" provides a new method for philosophy.

Next, Stefan Buijsman, Michael Klenk, and Jeroen van den Hoven dive into a subbranch of philosophy, ethics. In Chapter 3, they discuss the main ethical challenges raised by AI as a technology, as well as the potential methods to tackle those challenges. While they argue that ethical theories such as virtue ethics, consequentialism, and deontology are a helpful starting point, they believe these theories lack details for a more actionable and proactive "AI ethics." Instead, they propose that the best way forward is to consider design-approaches in the context of AI, such as "Design for Values," alongside interdisciplinary working methods. Their AI ethics overview paves the way for the next three chapters, which focus on a more specific ethical conundrum.

In Chapter 4, Laurens Naudts and Anton Vedder zoom in on the theme of AI and fairness. Taking as their point of departure one particular interpretation of fairness – namely fairness as non-arbitrariness – they analyze the distinction between procedural and substantive conceptions of fairness, as well as the relationship between fairness, justice, and equality. Subsequently, they distinguish distributive fairness approaches from socio-relational ones, and caution against the formalization of fairness by design as a form of techno-solutionism. Naudts and Vedder also emphasize that the design and regulation of fair AI systems is not an insular exercise, and that – beyond procedures and outcomes – sufficient attention must be paid to the social processes, structures, and relationships that inform and are co-shaped by the functioning of such systems.

Chapter 5 deals with another theme of ethical concern in the context of AI, namely moral responsibility. Lode Lauwaert and Ann-Katrien Oimann consider whether the use of autonomous AI causes a responsibility gap. After discussing how the notion of responsibility can be understood and what the responsibility gap is about, they explore in which ways it is sensible to assign responsibility to artificial systems and argue that their use does not necessarily lead to a responsibility gap. Moreover, they explain why, according to them, even if such a gap were to exist, it would not necessarily be problematic.

In the sixth and final chapter of this part, Gry Hasselbalch and Aimee Van Wynsberghe analyze the relationship between AI, power, and responsibility. They point out that AI has the potential to support solutions to counter sustainability concerns, while at the same time however also being unsustainable, given the high carbon emissions and the many ethical concerns it raises, from discrimination to surveillance and electoral micro-targeting. Making the plea that it is crucial to address the long-term sustainability of AI in light of its impact on our social, personal, and natural environments (also of future generations), they suggest a "sustainable" approach to AI. In Chapter 6, they hence argue that such an approach should

be inclusive in time and space, meaning that the past, present, and future of human societies, as well as the planet and environment, are considered equally important to protect and secure, including the integration of all countries in economic and social changes.

Part II: AI, Law and Policy

The second part of this book deals with the law and policy of AI, which constitute important tools to govern the technology's impact on society and its ethical challenges. In Chapter 7, Pierre Dewitte discusses AI's impact on privacy and its relationship with data protection law, arguing that the large-scale processing of personal data that AI systems enable also puts a strain on individuals' fundamental rights and freedoms. The chapter focuses in particular on the General Data Protection Regulation (GDPR) and describes its position and role within the broader European data protection regulatory framework. After introducing some of the GDPR's key concepts, it draws attention to certain tension points between the characteristics inherent to most AI systems and the general principles outlined in the GDPR, such as lawfulness, transparency, purpose limitation, data minimization, and accountability.

Chapter 8 deals with extra-contractual or tort liability in the context of AI, an area that is increasingly on legislators' radar given that the technology's use will inevitably lead to damage. Jan De Bruyne and Wannes Ooms discuss the main challenges that arise in this context and highlight that national law remains of great importance to tackle them. Focusing on the procedural elements of tort liability, including disclosure requirements and rebuttable presumptions, they also illustrate how existing tort law concepts are challenged by AI's characteristics, and which regulatory answers are available.

Chapter 9 deals with another legal domain that is impacted by AI, namely competition law. Friso Bostoen explains how companies increasingly rely on AI systems for (strategic) decisions, and how their use can have procompetitive effects, for instance, by facilitating the undercutting of competitors or improving recommendations. Yet he also cautions for AI's distortive effects on competition, for instance, when used to collude or to exclude competitors. He then analyzes to what extent such anticompetitive algorithmic practices are already covered by EU competition law by examining their use to conclude horizontal and vertical agreements, as well as to foster exclusionary and exploitative conduct.

In Chapter 10, Evelyne Terryn and Sylvia Martos Marquez move from competition law to consumer protection law, which traditionally focuses on protecting consumers' autonomy and self-determination – both of which are affected by the growing use of AI. In their analysis, they provide an overview of the most relevant consumer protection instruments in the EU legal order which apply to the context of AI. Finally, through a case study on dark patterns, they illustrate the shortcomings of the current consumer protection framework and argue for better safeguards.

Chapter 11, written by Jozefien Vanherpe, delves into the interface of AI and intellectual property law. She discusses the extent to which AI technology can be protected, whether it can be qualified as an author or inventor, and who holds ownership of AI-assisted and AI-generated output. She also considers how liability is allocated for intellectual property right infringements taking place by or through the intervention of an AI system and concludes that – despite the apparent enthusiasm for the use of AI in practice – there is also a hesitancy to provide additional incentive creation through (new or adapted) intellectual property legislation in the AI sphere.

In Chapter 12, the final chapter of this part, Karen Yeung and I provide a critical analysis of the European Union's AI Act. This regulation not only seeks to establish a single European market for AI, but is also meant to address some of the most pressing risks that AI systems pose to the health, safety, and human rights of individuals. We however question whether the AI Act can translate its noble aspirations into meaningful and effective protection for people whose lives are affected by AI systems. Through a critical examination of the proposed conceptual vehicles and regulatory architecture upon which the AI Act relies, we argue there are good reasons for skepticism, as many of the AI Act's provisions delegate critical regulatory tasks to AI providers, without adequate oversight or redress mechanisms.

Part III: AI across Sectors

Having looked at AI from a horizontal perspective in the previous two parts, Part III of this book focuses on a number of sectoral domains in which AI systems are used, and analyzes their more context-specific effects. In Chapter 13, Inge Molenaar, Duuk Baten, Imre Bárd, and Marthe Stevens discuss the implications of AI in the field of education. After introducing multiple existing perspectives on the role of AI in education, with an emphasis on an augmentation-approach that supports human strengths, they distinguish between students-faced, teacher-faced, and administrative AI solutions and trace how AI ethics in education was taken up in international and European policies. They close with an example of how intelligent innovations in the field of education can be cocreated in collaboration with educational professionals, scientists, and companies, drawing on the example of the "Dutch value compass for the digital transformation of education."

Chapter 14 turns to the permeation of AI in the media sector. Lidia Dutkiewicz, Noémie Krack, Aleksandra Kuczerawy, and Peggy Valcke first discuss the opportunities of the use of AI in media content gathering and production, media content distribution, fact-checking, and content moderation. They then zoom into some of the risks that arise in the context of AI-driven media applications, such as the lack of data availability, the lack of transparency, the adverse impact on the right to freedom of expression, as well as threats to media freedom and pluralism online, and threats to media independence. They also offer an overview of the EU legal framework that

aims to mitigate these risks, including the Digital Services Act, the European Media Freedom Act, and the AI Act.

In Chapter 15, Griet Verhenneman discusses the relationship between AI, healthcare data, and data protection law. She stresses that healthcare data are required not only for the research and development phases of AI but also for the establishment of evidence of compliance with legislation, such as the Medical Devices Regulation and the AI Act – which must occur without prejudice to other legal acts such as the GDPR. After introducing notions such as "real-world data," "evidence data," and "electronic health records," she discusses the role of healthcare data custodians and the impact of concepts like data ownership, patient autonomy, informed consent, and privacy and data protection-enhancing techniques in the context of AI healthcare applications.

Chapter 16, written by Katja Langenbucher, examines the role of AI in the financial world, where actors continuously process vast amounts of information, and increasingly do so with the aid of AI. To concretize the implications of this practice she describes AI scoring and creditworthiness assessments as an example of how AI systems are employed in financial services, which ethical challenges they raise, and how legal tools are balancing the advantages and challenges of this technology. Finally, she also looks ahead and cautions against AI-enabled scoring that ranges beyond the credit context, as it also extends toward people's social lives and facilitates novel forms of (unwarranted) control.

One area of increased control is the work sphere. In Chapter 17, Aída Ponce Del Castillo and Simon Taes provide an overview of the multifaceted aspects of AI and labor law, focusing on the profound questions arising in this intersection, from the impact on employment relationships, to the exercise of labor rights and social dialogue. After providing illustrations of common AI applications and discussing the use of automated decision-making and monitoring systems in the workplace, they also elucidate the most relevant rights and tools when it comes to the negotiation and implementation of AI in the workplace, as well as AI-related legislation with a work-oriented dimension.

Chapter 18, written by Rosamunde Van Brakel, introduces the use of AI in law enforcement and discusses the main legal, ethical, and social concerns this raises by focusing on one AI application in particular, namely predictive policing. In the last two decades, police forces in Europe and North America increasingly invested in such applications, of which she analyzes two types: predictive mapping and predictive identification. She discusses concerns around (the lack of information about) their effectiveness, as well as their impact on citizens and society.

In Chapter 19, I discuss the governance of algorithmic regulation in public administration – or the delegation of the application, execution, and enforcement of regulation to algorithmic systems. I contextualize public administrations' increased reliance on such digital technologies and discuss the ethical and legal conundrums that administrations face when outsourcing (part of) their tasks, from their impact on

the rule of law and digital sovereignty, to their discriminatory and intrusive effects. I also offer an overview of the legal framework that governs this practice in Europe, covering constitutional and administrative law, as well as data protection law and AI-specific law, all of which ought to be considered when public administrations seek to deploy algorithmic regulation.

Chapter 20 is concerned with the intersection of AI and armed conflicts. Katerina Yordanova reflects on the widespread development and adoption of AI and other digital technologies in warfare and recognizes the potential that AI carries for improving the applicability of the basic principles of international humanitarian law, if used in an accountable and responsible way. At the same time, she questions whether international humanitarian law is at all up to the task of addressing the threats posed by these technologies. After a description of the system, principles, and internal logic of international humanitarian law, she evaluates the role of AI systems in (non-)international armed conflicts and discusses some of the policy developments in this field, with the aim of contributing to the discussion on ex-ante regulation of AI systems for military purposes.

Finally, I close this book by offering some concluding remarks, drawing on the richness of the insights provided by the chapter authors and pointing to a few gaps that this book leaves unaddressed, which merit further research in the future.

OPEN QUESTIONS

To conclude this introduction, I would like to set out a few open questions that scholars in the field are often confronted with when it comes to the governance of AI, and that the authors of this book's chapters also had to deal with when writing their contributions.

A first question to ask is which human behavior in the context of AI should be subjected to (new or updated) binding legal rules, and which behavior can be left to non-legal norms. Not all ethical imperatives are also enshrined in legislation, nor are all legal rules necessarily reflecting an ethical norm. That said, law and ethics are strongly connected with each other, though neither can substitute the other,[10] and both have an important function in the AI governance context. In addition, laws are typically implemented through – though often also guided and indirectly shaped by – (government) policy, despite the fact that policy should ideally be no more than a "servant of the formal rule of law" to avoid excesses.[11] Yet what should the contours of the respective functions of *law*, *ethics*, and *policy* be? Which role can and should they play in reigning in the societal effects of the development and deployment of AI?

[10] See also Nathalie A. Smuha, "The EU approach to ethics guidelines for trustworthy artificial intelligence" (2019) *Computer Law Review International*, 2(4): 101.

[11] Theodore J. Lowi, "Law vs. public policy: A critical exploration" (2003) *Cornell Journal of Law and Public Policy*, 12(3): 501.