Introduction

Silja Voeneky, Philipp Kellmeyer, Oliver Mueller, and Wolfram Burgard

In the past decade, Artificial Intelligence (AI) as a general-purpose tool has become a disruptive force globally. By leveraging the power of artificial neural networks, deep learning frameworks can now translate text from hundreds of languages, enable real-time navigation for everyone, recognise pathological medical images, as well as enable many other applications across all sectors in society. However, the enormous potential for innovation and technological advances and the chances that AI systems provide come with hazards and risks that are not yet fully explored, let alone fully understood. One can stress the opportunities of AI systems to improve healthcare, especially in times of a pandemic, provide automated mobility, support the protection of the environment, protect our security, and otherwise support human welfare. Nevertheless, we must not neglect that AI systems can pose risks to individuals and societies; for example by disseminating biases, by undermining political deliberation, or by the development of autonomous weapons. This means that there is an urgent need for responsible governance of AI systems. This Handbook shall be a basis to spell out in more detail what could become relevant features of Responsible AI and how we can achieve and implement them at the regional, national, and international level. Hence, the aim of this Handbook is to address some of the most pressing philosophical, ethical, legal, and societal challenges posed by AI.

But mapping the uncertainties, benefits, and risks of AI systems in general and with regard to different sectors and assessing relevant ethical and legal rules requires a wide range of expertise from areas such as computer science, robotics, mathematical modelling, as well as trans- and interdisciplinary normative analyses from law, philosophy, and ethics by authors from different continents. Therefore, the authors of this Handbook explore the technical and conceptual foundations as well as normative aspects of Responsible AI from many different angles.

OUTLINE

The Handbook consists of eight parts and begins with the foundations of Responsible AI (first part), the current and future approaches to AI governance (second part) not limited to European and US approaches. Authors further analyse liability schemes (third part) and shed light on the core problem of fairness and non-discrimination in AI systems (fourth part) before approaches in responsible data governance are examined in more detail (fifth part). The authors of the sixth and seventh parts discuss justified governance approaches of specific sectors of AI systems: these systems can be part of such important fields as corporate governance, including financial

https://doi.org/10.1017/9781009207898.001 Published online by Cambridge University Press

services, and the healthcare sector, including neurotechnology. Authors of the eighth part tackle especially problematic and challenging issues such as the use of AI for security applications and in armed conflict.

PART I: FOUNDATIONS

The first chapter, written by *Wolfram Burgard*, describes features of AI systems, introducing *inter alia* the notions of machine learning and deep learning as well as the use of AI systems as part of robotics. In doing so, *Burgard* provides an overview of the technological state of the art from the perspective of robotics and computer science.

Jaan Tallinn and *Richard Ngo* propose an answer to the question of what Responsible AI means by offering a framework for the deployment of AI which focuses on two concepts: delegation and supervision. This framework aims towards building 'delegate AIs' which lack goals of their own but can perform any task delegated to them. However, AIs trained with hardcoded reward functions, or even human feedback, often learn to game their reward signal instead of accomplishing their intended tasks. Thus, *Tallinn* and *Ngo* argue that it will be important to develop more advanced techniques for continuous, high-quality supervision; for example, by evaluating the reasons which AI-systems give for their choices of actions. These supervision techniques might be made scalable by training AI-systems to generate reward signals for more advanced AI-systems.

After this, the philosopher *Catrin Misselhorn* provides an overview of the core debates in artificial morality and machine ethics. She argues that artificial moral agents are AI systems which are able to recognise morally relevant factors and take them into account in their decisions and actions. *Misselhorn* shows that artificial morality is not just a matter of science fiction scenarios but rather an issue that has to be considered today. She provides a basis for what Responsible AI means by laying down conceptual foundations of artificial morality and discussing related ethical issues.

The philosopher *Johanna Thoma* focuses, as part of the fourth chapter, on the 'moral proxy problem,' which arises when an autonomous artificial agent makes a decision as a proxy for a human agent without it being clear for whom specifically it does so. *Thoma* identifies two categories of agents an artificial agent can be a proxy for: *low-level* agents (individual users or the kinds of human agents that are typically replaced by artificial agents) and *high-level* agents (designers, distributors, or regulators). She shows that we do not get the same recommendations under different agential frames: whilst the former suggests the agents be programmed without risk neutrality, which is common in choices made by humans, the latter suggests the contrary, since the choices are considered part of an aggregate of many similar choices.

In the final chapter of the first part, the philosopher *Christoph Durt* develops a novel view on AI and its relation to humans. *Durt* contends that AI is neither merely a tool, nor an artificial subject, nor necessarily a simulation of human intelligence, but rather AI is defined by its *interrelational character*. According to this author, misconceptions of AI have led to far-reaching misunderstandings of the opportunities and risks of AI. Hence, a more comprehensive concept of AI is needed to better understand the possibilities of Responsible AI. He argues that the setup of the *Turing Test* is already deceptive, as this test avoids difficult philosophical questions by passing on the burden to an evaluator, and delineates a more comprehensive picture according to which AI integrates into the human lifeworld through its interrelations with humans and data.

https://doi.org/10.1017/9781009207898.001 Published online by Cambridge University Press

PART II: APPROACHES TO AI GOVERNANCE

The second part of the Handbook, which focuses on current and future approaches to AI governance, begins with a chapter by the philosopher *Mathias Risse*. He reflects on the medium- and long-term prospects and challenges democracy faces from AI. *Risse* argues that both technologists and citizens need to engage with ethics and political thoughts in order to build and maintain a democracy-enhancing AI infrastructure, and stresses the need for Responsible AI as he points out AI's potential to greatly strengthen democracy, but only with the right efforts. His answer starts with a critical examination of the relation between democracy and technology with a historical perspective before outlining a 'techno skepticism' prevalent in several grand narratives of AI. *Risse* explores the possibilities and challenges that AI may lead to and argues that technology critically bears on what forms of human life get realised or imagined. According to the author, AI changes the materiality of democracy by altering how collective decision making unfolds and what its human participants are like.

Next, *Thomas Burri*, an international law scholar, examines how general ethical norms on AI diffuse into domestic law without engaging international law. *Burri* discusses various frameworks for 'ethical AI' and shows how they influenced the European Union Commission's proposal for the draft 2021 AI Act. Hence, this chapter reveals the origins of the EU proposal and explains the substance of the future EU AI regulation.

From an interdisciplinary angle, the mathematician *Thorsten Schmidt* and the ethics and international law scholar *Silja Voeneky* propose a new adaptive regulation scheme for AI-driven high-risk products and services as part of a future Responsible AI governance regime. They argue that current regulatory approaches with regard to these products and services, including the draft 2021 EU AI Act, have to be supplemented by a new regulatory approach. At its core, the adaptive AI regulation proposed by the authors requires that private actors, like companies developing and selling high-risk AI-driven products and services, pay a proportionate amount of money as a financial guarantee into a fund before the AI-based high-risk product or service enters the market. *Schmidt* and *Voeneky* spell out in more detail what amount of regulatory capital can be seen as proportionate and what kind of accompanying rules are necessary to implement this adaptive AI regulation.

In chapter nine, the law scholars *Weixing Shen* and *Yun Liu* focus on China's AI regulation. They show that there is no unified AI law today in China but argue that many provisions from Chinese data protection law are in part applicable to AI systems. The authors particularly analyse the rights and obligations from the Chinese Data Security Law, the Chinese Civil Code, the E-Commerce Law, and the Personal Information Protection Law; they finally introduce the Draft Regulation on Internet Information Service Based on Algorithm Recommendation Technology and explain the relevance of these regulations with regard to algorithm governance.

Current and future approaches to AI governance have to be looked at from a philosophical perspective, too, and *Thomas Metzinger* lists the main problem domains related to AI systems from a philosophical angle. For each problem domain, he proposes several measures which should be taken into account. He starts by arguing that there should be worldwide safety standards concerning the research and development of AI. Additionally, a possible AI arms race must be prevented as early as possible. Thirdly, he stresses that any creation of artificial consciousness should be avoided, as it is highly problematic from an ethical point of view. Besides, he argues that synthetic phenomenology could lead to non-biological forms of suffering and might lead to a vast increase of suffering in the universe, as AI can be copied rapidly, and that in the field of AI systems there is the risk of unknown risks.

In the final chapter of this part, the law and technology scholar *Jonathan Zittrain* begins with the argument that there are benefits to utilising scientific solutions discovered through trial and error rather than rigorous proof (though aspirin was discovered in the late nineteenth century, it was not until the late twentieth century that scientists were able to explain how it worked), but argues that doing so accrues 'intellectual debt'. This intellectual debt is compounding quickly in the realm of AI, especially in the subfield of machine learning. Society's movement from basic science towards applied technology that bypasses rigorous investigative research inches us closer to a world in which we are reliant on an oracle AI, one in which we trust regardless of our ability to audit its trustworthiness. *Zittrain* concludes that we must create an intellectual debt 'balance sheet' by allowing academics to scrutinise the systems.

PART III: RESPONSIBLE AI LIABILITY SCHEMES

The next part of the Handbook focuses on Responsible AI liability schemes from a legal perspective. First of all, *Christiane Wendehorst* analyses the different potential risks posed by AI as part of two main categories, safety risks and fundamental rights risks, and considers why AI challenges existing liability regimes. She highlights the fact that liability for fundamental rights risks is largely unchartered while being AI-specific. *Wendehorst* argues that a number of changes have to be made for the emerging AI safety regime to be used as a 'backbone' for the future AI liability regime if this is going to help address liability for fundamental rights risks. As a result, she suggests that further negotiations about the AI Act proposed by the European Commission should be closely aligned with the preparatory work on a future AI liability regime.

Secondly, the legal scholar *Jan von Hein* analyses and evaluates the European Parliament's proposal on a civil liability regime for AI against the background of the already existing European regulatory framework on private international law, in particular the Rome I and II Regulations. He argues that the draft regulation proposed by the European Parliament is noteworthy from a private international law perspective because it introduces new conflicts rules for AI. In this regard, the proposed regulation distinguishes between a rule delineating the spatial scope of its autonomous rules on strict liability for high-risk AI systems, on the one hand, and a rule on the law applicable to fault-based liability for low-risk systems, on the other hand. *Von Hein* concludes that, compared with the current Rome II Regulation, the draft regulation would be a regrettable step backwards.

PART IV: FAIRNESS AND NON-DISCRIMINATION

The fourth part of the Handbook, which links fairness and non-discrimination in AI systems to the concept of Responsible AI, begins with a chapter by the philosopher *Wilfried Hinsch*. He focuses on statistical discrimination by means of computational profiling. He argues that because AI systems do not rely on human stereotypes or rather limited data, computational profiling may be a better safeguard of fairness than humans. He starts by defining statistical profiling as an estimate of what individuals will do by considering the group of people they can be assigned to, and explores which criteria of fairness and justice are appropriate for the assessment of computational profiling. *Hinsch* argues that discrimination constitutes a rule-guided social practice that imposes unreasonable burdens on specific people. He spells out that even statistically correct profiles can be unacceptable considering reasons of procedural fairness or substantive justice. Because of this, he suggests a fairness index for profiles to determine procedural fairness.

In Chapter 15, the legal scholar *Antje von Ungern-Sternberg* focuses on the legality of discriminatory AI, which is increasingly used to assess people (profiling). As many studies show that the use of AI can lead to discriminatory outcomes, she aims to answer the question of whether the law as it stands prohibits objectionable forms of differential treatment and detrimental impact. She takes up the claim that we need a 'right to reasonable inferences' with respect to discriminatory AI and argues that such a right already exists in antidiscrimination law. Also, *von Ungern-Sternberg* shows that the need to justify differential treatment and detrimental impact implies that profiling methods correspond to certain standards, and that these methodological standards have yet to be developed.

PART V: RESPONSIBLE DATA GOVERNANCE

The fifth part of the Handbook analyses problems of responsible data governance. The legal scholar *Ralf Poscher* sets out to show, in Chapter 16, how AI challenges the traditional understanding of the right to data protection and presents an outline of an alternative conception that better deals with emerging AI technologies. He argues that we have to step back from the idea that each and every instance of personal data processing concerns a fundamental right. For this, *Poscher* explains how the traditional conceptualisation of data protection as an independent fundamental right on its own collides with AI's technological development, given that AI systems do not provide the kind of transparency required by the traditional approach. And secondly, he proposes an alternative model, a no-right thesis, which shifts the focus from data protection as an independent right to other existing fundamental rights, such as liberty and equality.

The legal scholar *Boris Paal* also identifies a conflict between two objectives pursued by data protection law, the comprehensive protection of privacy and personal rights and the facilitation of an effective and competitive data economy. Focusing on the European Union's General Data Protection Regulation (GDPR), the author recognises its failure to address the implications of AI, the development of which depends on the access to large amounts of data. In general, he argues, that the main principles of the GDPR seem to be in direct conflict with the functioning and underlying mechanisms of AI applications, which evidently were not considered sufficiently whilst the regulation was being drafted. Hence, *Paal* argues that establishing a separate legal basis governing the permissibility of processing operations using AI-based applications should be considered.

In the last chapter of the fifth part, the legal scholars *Sangchul Park*, *Yong Lim*, and *Haksoo Ko*, analyse how South Korea has been dealing with the COVID-19 pandemic and its legal consequences. Instead of enforcing strict lockdowns, South Korea imposed a robust, AI-based contact tracing scheme. The chapter provides an overview of the legal framework and the technology which allowed the employment of this technology-based contact tracing scheme. The authors argue that South Korea has a rather stringent data-protection regime, which proved to be the biggest hurdle in implementing the contact tracing scheme. However, the state introduced a separate legal framework for extensive contact tracing in 2015, which was reactivated and provided government agencies with extensive authority to process personal data for epidemiological purposes.

PART VI: RESPONSIBLE CORPORATE GOVERNANCE OF AI SYSTEMS

The sixth part looks at responsible corporate governance of AI systems and it starts by exploring the changes that AI brings about in corporate law and corporate governance. The legal scholar

https://doi.org/10.1017/9781009207898.001 Published online by Cambridge University Press

Jan Lieder argues that whilst there is the potential to enhance the current system, there are also risks of destabilisation. Although algorithms are already being used in the board room, lawmakers should not consider legally recognizing e-persons as directors and managers. Rather, scholars should evaluate the effects of AI on the corporate duties of boards and their liabilities. Finally, *Lieder* suggests the need for transparency in a company's practices regarding AI for awareness-raising and the enhancement of overall algorithm governance, as well as the need for boards to report on their overall AI strategy and ethical guidelines relating to the responsibilities, competencies, and protective measures they established.

In Chapter 20, it is shown how enforcement paradigms that hinge on descriptions of the inner sphere and conduct of human beings may collapse when applied to the effects precipitated by independent AI-based computer agents. It aims to serve as a conceptual sketch for the intricacies involved in autonomous algorithmic collusion, including the notion of concerted practices for cases that would otherwise elude the cartel prohibition. *Stefan Thomas*, a legal scholar, starts by assessing how algorithms can influence competition in markets before dealing with the traditional criteria of distinction between explicit and tacit collusion. This might reveal a potential gap in the existing legal framework regarding algorithmic collusion. Finally, *Thomas* analyses whether the existing cartel prohibition can be construed in a manner that captures the phenomenon appropriately.

Matthias Paul explores in the next chapter the topic of AI systems in the financial sector. After outlining different areas of AI application and different regulatory regimes relevant to robo-finance, he analyses the risks emerging from AI applications in the financial industry. The author argues that AI systems applied in this sector usually do not create new risks. Instead, existing risks can actually be mitigated through AI applications. *Paul* analyses personal responsibility frameworks that have been suggested by scholars in the field of robo-finance, and shows why they are not a sufficient approach for regulation. He concludes by discussing the draft 2021 EU AI Act as a suitable regulatory approach based on the risks linked to specific AI systems and AI-based practices.

PART VII: RESPONSIBLE AI IN HEALTHCARE AND NEUROTECHNOLOGY

As another important AI-driven sector, the seventh part of the Handbook focuses on Responsible AI in healthcare and neurotechnology governance. The legal scholars *Fruzsina Molnár-Gábor* and *Johanne Giesecke* begin by setting out the key elements of medical AI. They consider the aspects of how the application of AI-based systems in medical contexts may be guided according to international standards. The authors argue that among the frameworks that exist, the World Medical Association's codes appear particularly promising as a guide for standardisation processes. *Molnár-Gábor* and *Giesecke* sketch out the potential applications of AI and its effects on the doctor–patient relationship in terms of information, consent, diagnosis, treatment, aftercare, and education.

In Chapter 23, the legal scholar *Christoph Krönke* focuses on the legal challenges healthcare AI Alter Egos face, especially in the EU, as these AI Alter Egos have two main functions, collecting a substantive database and proposing diagnoses. The author spells out the relevance of European data protection laws and analyses the European Medical Devices Regulation (MDR) with regard to the responsible governance of these entities. *Krönke* argues that AI Alter Egos are regulated by an appropriate legal framework in the EU today, but it nevertheless has to be open for developments in order to remain appropriate.

In the following chapter, neurologist and neuroethics scholar *Philipp Kellmeyer* sets out a human-rights based approach for governing AI-based neurotechnologies. *Kellmeyer* outlines the

current scholarly discussion and policy initiatives about neurorights and discusses how to protect mental privacy and mental integrity. He argues that mental privacy and integrity are important anthropological goods that need to be protected from unjustified interferences. He argues that while existing human rights provide a sufficient legal basis, an approach is required that makes these rights actionable and justiciable to protect mental privacy and mental integrity.

In the final chapter of this part, the philosophers *Boris Essmann* and *Oliver Mueller* address AI-supported neurotechnology, especially Brain–Computer Interfaces (BCIs) that may in the future supplement and restore functioning in agency-limited individuals or even augment or enhance capacities for natural agency. The authors propose a normative framework for evaluating neurotechnological and AI-assisted agency based on 'cyberbilities' that can be part of a Responsible AI framework. 'Cyberbilities' are capabilities that emerge from human–machine interactions in which agency is distributed across human and artificial elements. *Essmann* and *Mueller* suggest a list of 'cyberbilities' that is meant to support the well-being of individuals.

PART VIII: RESPONSIBLE AI FOR SECURITY APPLICATIONS AND IN ARMED CONFLICT

The eighth and final part of this Handbook discusses the highly controversial use of Responsible AI for security applications and in armed conflict. The legal scholar *Ebrahim Afsah* outlines different implications of AI for the area of national security. He argues that while AI overlaps with many challenges to the national security arising from cyberspace, it also creates new risks, including the development of autonomous weapons, the enhancement of existing military capabilities, and threats to foreign relations and economic stability. Most of these risks, however, *Afsah* argues, can be subsumed under existing normative frameworks.

In the next chapter, the political philosopher *Alex Leveringhaus* spells out ethical concerns regarding autonomous weapons systems (AWS) by asking whether lethal autonomous weapons are morally repugnant and whether this entails that they should be prohibited by international law. *Leveringhaus* surveys three prominent ethical arguments against AWS: firstly, autonomous weapons systems create 'responsibility gaps'; secondly, that their use is incompatible with human dignity; and, thirdly, that autonomous weapons systems replace human agency with artificial agency. He argues that some of these arguments fail to show that autonomous weapons systems are morally different from more established weapons. However, the author concludes that autonomous weapons systems are problematic due to their lack of predictability.

In the final chapter of this Handbook, the legal scholar *Dustin Lewis* discusses the use of Responsible AI during armed conflict. The scope of this chapter is not limited to lethal autonomous weapons but also encompasses other AI-related tools and techniques related to warfighting, detention, and humanitarian services. For this, he explores the requirements of international law and outlines some preconditions necessary to respect international law. According to *Lewis*, current international law essentially presupposes humans as legal agents. From that premise, the author argues that any employment of AI-related tools or techniques in an armed conflict needs to be susceptible to being administered, discerned, attributed, understood, and assessed by human agents.

After this outline about the core issues discussed with regard to the concept of Responsible AI, we want to note that many chapters of the Handbook have strong links to an international and interdisciplinary virtual research conference on "*Global Perspectives on Responsible AI*". We convened this conference in June 2020 based at the *Freiburg Institute for Advanced Studies* and edited this Handbook in order to shed more light on the transformations that are based on the rise

https://doi.org/10.1017/9781009207898.001 Published online by Cambridge University Press

of AI systems, their impact on our societies, and the challenges for the responsible governance and regulation of AI. Although the chapters of this Handbook shall not – and cannot – answer all questions with regard to AI governance and more problems have to be discussed and solved in the forthcoming years, we as editors agree that Responsible AI governance shall be conducive to scientific and technological progress, to our stability and flourishing as individuals and as humanity. We hope that the perspectives, analyses, and proposals for a concept of Responsible AI in this Handbook will provide a basis for fostering deliberations to develop and spell out proportional approaches to AI governance and enable scholars, scientists, and other actors to discuss normative frameworks for AI that allow societies, states, and the international community to unlock the potential for responsible innovation in this important field.

https://doi.org/10.1017/9781009207898.001 Published online by Cambridge University Press