AI in the EU: Ethical Guidelines as a Governance Tool

Larsson, Stefan

Published in:
The European Union and the Technology Shift

DOI:
10.1007/978-3-030-63672-2_4

2021

Document Version:
Publisher's PDF, also known as Version of record

CHAPTER 4

AI in the EU: Ethical Guidelines as a Governance Tool

Stefan Larsson

Introduction: Why Ethics Guidelines?

In socio-legal research there is a long tradition of studying law's relationship to society, for example, its use as an instrument of social control (Cotterrell 1992; Fuller 1975), or how also informal rules may govern (Ellickson 1994). Much comment has been made on the interplay between 'soft' and 'hard' law (cf. Abbott and Snidal 2000). Particularly emerging technologies, it seems, provide for an area of significant regulatory sensitivity in terms of striking a balance between promises of innovation, on the one hand, and concerns about risks and a related lack of public confidence, on the other (Mandel 2009). Artificial intelligence (AI), correspondingly, is currently a technologically driven field in the midst of such a governance challenge, not the least in Europe (Larsson 2020).

The 2010s have seen significant progress in the field of AI, especially within the framework of machine learning. Partly to promote this

S. Larsson (✉)
Department of Technology and Society, Lund University, Lund, Sweden
e-mail: stefan.larsson@lth.lu.se

A. Bakardjieva Engelbrekt et al. (eds.), *The European Union and the Technology Shift*, https://doi.org/10.1007/978-3-030-63672-2_4

development, in April 2018, the European Commission adopted an AI strategy. The Commission appointed 52 experts to a High-Level Expert Group on Artificial Intelligence (AI HLEG) tasked with making policy and investment recommendations and offering guidance on ethical issues related to the use of AI in Europe. In April 2019, AI HLEG published the document *Ethics Guidelines for Trustworthy AI*, which—despite explicitly stating that the guidelines are 'not meant to provide legal advice or to offer guidance on compliance with applicable laws'—clearly places issues of accountability, human centricity, transparency and privacy at the heart of the development of trustworthy AI (HLEG 2019b). Far from being an isolated phenomenon, AI HLEG's guidelines can be viewed as part of a recent trend towards the development of a host of ethical guidelines by corporations, researcher groups and government agencies. Although many of these overlap with existing legislation, it is often unclear exactly how legislation and guidelines are intended to interact. In particular, there is often ambiguity concerning the intended implementation of ethical principles. Guidelines tend to focus on aspirational needs at a general level but are often procedurally weak.

On a more general level, the ethical guidelines for AI prepared by the European Commission's expert group are a clear sign of an ongoing governance challenge facing the European Union and its member states. As an indication, during her candidature, President of the European Commission Ursula von der Leyen stated that 'in my first 100 days in office, I will put forward legislation for a coordinated European approach on the human and ethical implications of Artificial Intelligence' (von der Leyen 2019: 13). Consequently, in February 2020, the European Commission published a White Paper on AI, addressing the risks associated with its use, and including policy and regulatory options 'towards an ecosystem for excellence and trust' (European Commission 2020). Much as in earlier governance of emerging technologies (cf. Mandel 2009), part of the challenge lies in balancing regulation with the belief that exists in technical innovation and with those general social developments that AI and its methods can contribute to, which one would not wish to hamper with excessive regulation.

This chapter focuses on ethical guidelines as a tool for AI governance, in order to highlight the interaction between guidelines and law and to discuss why the specific nature of AI development has led to ethical issues in particular taking such a prominent place. A wider discussion is underway regarding the governance of AI that this chapter seeks to contribute

to by providing a European perspective. The chapter is structured around the following three questions:

1. How can we understand a data-dependent AI based on its interaction with human values and societal structures?
2. Why do contemporary notions of AI governance so often come in the form of ethical guidelines?
3. Which ethical principles are most commonly highlighted as being key to AI governance, and why?

The first of these questions is preceded by the fact that the very definition of AI is a matter for debate, depending on which field or discipline the person defining the term works in. Here, I argue for the need to view these technologies in their applied context and in interaction with human values and societal structures, something not least underlined by machine learning's dependence on large amounts of sample data on which to base its computations (for a discussion on the 'normative mirroring effect', see Larsson 2019). An AI system can therefore reproduce not only positive and intended expressions and structures but also the more problematic sides and patterns of human behaviour, for example, gender inequality and other forms of discrimination. Powerful technologies also carry inherent and obvious risks of various types of malicious misuse. *The second* question can be placed in a broader field of research into instruments of control with various purposes but that often share principles and key values such as control over data, reasonable levels of transparency and the division of accountability. The field of AI therefore follows many paths, from various governance functions and instruments of control in the form of legislation, standardisation and, not least, ethical guidelines. *The third* question is twofold in as much as it both highlights the types of AI-related challenges that receive most attention and points out the emerging insight into the societal challenges that therefore form the basis of the ethical guidelines. One important aspect is the temporal gap between a slow and somewhat drawn-out, but democratic and politically supported, legislative process and the rapid development that characterises AI and its contemporary data-dependent applications (cf. Larsson 2020).

The following subsection "From Area of Research to Regulatory Concept" addresses the definitional fuzziness of the AI concept. The importance of such a conceptual focus lies in the functional difference between AI as a rich research area with a long prehistory and AI as a

normative concept found in guidelines meant to regulate human activities. Subsection "What Is It That Needs to be Governed?" addresses in more detail what it is that needs to be governed, based on growing insight within critical AI research concerning consequences of applied AI, especially in terms of discriminatory or skewed outcomes of algorithmic processes. Both unpredictability and the challenges of explaining how algorithms and AI-systems achieve a given result, outcome or solution to a specific problem have led to the emergence of transparency as a key issue. Thereafter, in subsection "What Does AI Governance Entail? Keeping Society-in-the-Loop", the term AI governance and the concept of controlling AI are addressed specifically. This is done based on insights into the ongoing interplay between society and AI, focusing both on the need to constantly evaluate applied AI based on society's norms and ethics and on how today's AI systems are often dependent on large amounts of training data in order to be able to solve problems. In many cases, these data include images of people or a quantification of human expression and social structures, meaning that applied AI systems are interacting with various areas of society. Here, the term *society-in-the-loop* is used in order to demonstrate how innovation must reciprocate society's expectations and needs. This also implies a need to develop arguments about why AI needs to be controlled. In the next subsection "Ethics Guidelines", the concept of governance is addressed, with an emphasis on ethical guidelines—the kernel of this chapter. The European perspective is studied in some depth by analysing the key guidelines thus far produced, that is, *Ethics Guidelines for Trustworthy AI*, and here it is pointed to the governance issues of two particularly important but convoluted concepts (in subsection "An Ecosystem of Trust? Risk and Transparency"): risk and transparency. In the concluding subsection "Summing up: From Principles to Effective Implementation", recommendations and possible ways forward are presented, with the intention of pointing out development areas and relevant issues for legislators, public authorities and those researching, developing and deploying AI systems.

From Area of Research to Regulatory Concept

Despite the attention lavished on AI and machine learning in both the media and European policy-making, there is no broad consensus regarding how best to define AI. A number of definitions have been launched in both research papers and government agency reports, but a major

challenge lies in the dynamic nature of this evolving field. Here, I would like to highlight the dynamics of the conceptual framework as it has been discussed within traditional AI research: firstly, by presenting some discernible key elements and then by demonstrating which aspects AI HLEG considers to be important. Finally, in the light of the challenges presented by AI in its application and interaction with societal values and structures, I argue that, from a multidisciplinary perspective, there is a need for a reconceptualisation of AI when dealing with issues of governance. Definition in itself is a form of conceptual control that has an impact on regulatory debate (cf. Larsson 2017), which is why caution should be exercised when developing definitions of such a multifaceted concept as AI. This is of particular significance as the concept of AI, describing a rich and dynamic research tradition, enters into a regulatory discourse to become a concept of key importance for governance of human, corporate and governmental activities.

In conjunction with the publication of AI HLEG's *Ethics Guidelines for Trustworthy AI* (HLEG 2019b), a document was published, titled *A Definition of AI: Main Capabilities and Disciplines* (HLEG 2019a), with the intention to clarify certain aspects of AI as a scientific discipline and technology. The stated purposes of the document were 'to avoid misunderstandings, to achieve a shared common knowledge of AI that can be fruitfully used also by non-AI experts, and to provide useful details that can be used in the discussion on both the AI ethics guidelines and the AI policies recommendations' (2019a: 1). AI HLEG took as its point of departure the definition of AI proposed in the European Commission's Communication on *Artificial Intelligence for Europe* (European Commission 2018), published in April 2018, which it has subsequently developed. The European Commission's communication defines AI thus:

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.
>
> AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).

This definition is largely focused on autonomy—that is, that there is an element of agency in AI-based systems—and points out that these systems can either be embedded in physical devices or exist purely as software. At the same time, these examples provide an indication of AI HLEG's thinking and, in the long run, which objects their ethical guidelines are intended to control. In the category of software-based AI systems, they mention voice assistants, image analysis software, search engines and speech and face recognition systems. In the category of physical, hardware-based AI systems, they identify applications such as advanced robots, autonomous cars, drones and devices connected to the Internet of things (IoT). As they emphasise autonomy, the guidelines may be interpreted as not applying to certain drones or connected devices, but only to those with an autonomous or self-learning component. The question of what constitutes 'advanced robots' cannot necessarily be answered by a simple line of demarcation. This ties into one significant problem in defining AI—a problem linked to the technology's dynamic side; to a certain extent, this definition of AI encompasses something as yet unachieved. A report from Stanford University written by the AI100 Standing Committee and Study Panel refers to this phenomenon as the 'AI effect' or the 'odd paradox': once AI-based technology enters into the public domain, it is no longer considered to be AI (Stone et al. 2016). In much the same way as 'advanced robots' in 2020 are not likely describing the same phenomena as they were at the beginning of the 1990s, and nor will they be in 2030. So, the conceptual line of demarcation distinguishing AI changes in line with what is technologically possible and how these methods become available for wider use. AI, to some extent, seems to be a concept reserved for the publicly unattainable.

As AI HLEG notes, intelligence itself, although explicitly referenced in the term AI, is a vague concept that has been attached to the technology since the founding of the field of research. There may be at least over 70 definitions of the term intelligence (Legg and Hutter 2007) focusing on various constituent aspects, with different emphases on problem solving, improvement and learning over time, good performance in complex environments or generalisability in being able to solve various types of problems without specific training in each individual type of problem domain. The latter is often described as part of the quest to achieve a general intelligence that—unlike today's more narrow, limited-domain AI—is able to achieve an overarching intelligence capable of solving various types of complex problems. Here, dynamic human intelligence is often considered

as an example; however, the concept of intelligence also has a number of associations to human abilities other than problem solving, such as emotional intelligence and self-awareness.

One can therefore say that, at the beginning of the 2020s, AI is an umbrella term covering a number of different technologies and analytical methods such as machine learning, natural language processing, image recognition, neural networks and deep learning. Machine learning in particular can be highlighted as, in simple terms at least, dealing with methods for allowing computers to 'learn' based on data alone, without having been programmed for a specific task. Machine learning is a field that has developed extremely strongly during the late 2010s as a result of access to historically unparalleled quantities of available digital data and greatly increased analytical processing power. The term machine learning was coined in 1959 by pioneering researcher Arthur Samuel, who created one of the world's first gaming computers (Samuel 1959). Since then, however, the field has gone from being a sub-discipline of AI, the main goal of which was to strive for artificial intelligence, to becoming a more practically oriented field of research focused on prediction based on training data. These days, while the field is generally included in AI, it is also closely linked to statistics and image recognition, where machine learning has proved to be extremely useful. Key to machine learning specifically, as well as AI in general, is, of course, the algorithms that are used, developed and studied to create learning effects in software and provide probability assessments.

The complexity of the conceptual framework led AI HLEG to advance a somewhat multifaceted definition that expands on the European Commission's original definition of AI. This expanded definition also encompasses AI functionality in its systematic context (i.e. that it is often part of a greater whole), machine learning's division between structured and unstructured data and the fact that AI systems are mainly goal-oriented to achieve something defined by a human being:

> Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt

> their behaviour by analysing how the environment is affected by their previous actions. (AI HLEG 2019a: 6)

There are various aspects of AI that must be captured in any definition of the phenomenon, the most important to current AI development and deployment tending to revolve around: (a) autonomy/agency, (b) self-learning from large quantities of sample data and (c) the level of generalisable learning. It is sometimes said that, despite the rapid pace of development in the field, we are working within the framework of a weak or narrow intelligence that remains limited to narrowly defined problem areas. In some areas of research at least, efforts are being made to develop more general intelligent applications that would be able to transfer insights from one specific domain to another.

Despite AI HLEG's *Ethics Guidelines for Trustworthy AI* largely dealing with ethical, legal, social and humanistic issues, these subject areas are notable by their absence when AI is described as a research discipline.

> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

The lack of references to the humanities or social-scientific disciplines in this definition is noteworthy, especially considering that the document in question highlights the challenges presented by explainable AI and biased or skewed data. In other words, there is a conceptual dichotomy between (i) AI HLEG's definition of AI research as a largely mathematics-based data or software-oriented research discipline and (ii) the need for AI-related research in the fields of jurisprudence, the humanities and social sciences. That said, it is possible that it is only a matter of time until these fields of research converge more clearly and their various scientific backgrounds are reflected in the definition of AI research.

Lastly, the description and understanding of AI—its meaning, concept and metaphors—must be managed with a certain amount of sensitivity to how these technologies, in their use and development, interact with society. This reasonably implies that one should view the privilege of definition

as an important part of, or precursor to, the development of the regulation and governance of AI. In other words, there is reason to complement mathematical and computer science-based definitions of AI in the step towards becoming a regulatory concept, especially when one's purpose is to better understand the implications of the social application of these technologies and methods and the consequent need for regulation.

What Is It That Needs to Be Governed?

Whether we refer to it as self-learning technology or autonomous systems, AI is often described as having enormous potential in terms of offering customised and relevant services, improved preventive diagnostics and automated decision-making in such diverse domains as healthcare, the public sector and self-driving vehicles. This development has been rapid in terms of machine-learning capacity, with technologies such as neural networks, deep learning and generative adversarial networks (GANs) that can generate synthetic data to facilitate the creation of realistic (but fake) images (cf. de Vries 2020). This has enormous potential in a range of data-dependent fields such as retail, healthcare and public administration. At the same time, insight is growing regarding the ethical and normative challenges presented by applied AI. One focus of modern AI development is on learning itself, that is, that the underlying models are adapted and modified based on the data (the examples) presented. A prediction, diagnosis or individual adaption will therefore reflect or reproduce the underlying data. As AI becomes part of everyday life—in our social media feeds, playlist recommendations and credit ratings—social structures and individual behaviour and values will provide an enormous amount of data to be collected and processed. This carries the risk that the undesirable biases and inequities inherent in society will be reproduced in AI services. At the same time, the complexity of the technology and sheer volume of data make these processes opaque and difficult to inspect, hence the comparison with a black box (cf. de Laat 2018; Lepri et al. 2018) having market-based and societal implications (Pasquale 2015).

If we examine possible distorting effects more closely, studies have for commercial systems found that there are AI systems that have demonstrably lower precision for women and people of colour (Buolamwini and Gebru 2018) and found predictive bias in the performance of pedestrians with different skin tones (Wilson et al. 2019). One of the most debated cases regards the so-called COMPAS system designed for use in the US

court system to assess the risk of recidivism, but was revealed to be racially biased (Pro Publica 2016). A further example is gender-discriminating job-recommendation systems that automatically recommend jobs with higher salaries to men (Datta et al. 2015). It has also been demonstrated that widely used image databases like ImageNet, with uneven cultural, gender and ethnic distribution, have unwanted effects on learning algorithms (Shankar et al. 2017). As a result, cultural attributes, such as wedding dresses, were classified with much less precision for cultures not well represented in the underlying database, that was biased towards North American and European attributes. This led the researchers advocating for broader geo-representation in training databases for machine learning (Shankar et al. 2017). Normative and ethical implications of studies like that mentioned are providing with arguments for AI governance, as these learning technologies become active applications interacting with social values and societal structures.

What Does AI Governance Entail? Keeping Society-in-the-Loop

There are a number of ways to understand governance as it relates to AI. A report from Oxford University's Future of Humanity Institute focuses on the extreme risks from advanced AI. The Institute is led by Professor of Philosophy Nick Bostrom, who is most renowned for his research into superintelligence; the idea that machine intelligence will achieve such general intelligence that it will become the dominant life form on the planet, posing an existential threat to humanity (cf. Bostrom 2014). There are, however, more mundane implications of applied AI that have led to contemporary needs for governance. When assessing the forms of 'soft law', one, however, also needs to address the relationship to already present 'hard law'. Some areas of AI development and application are undoubtedly already regulated, for example, with regard to privacy in the form covered by the General Data Protection Regulation (GDPR). Nevertheless, the field of governance is characterised by a great deal of activity in the area of ethical guidelines developed to influence the development and application of AI, especially in terms of dealing with those practices deemed especially problematic. Before an examination of specific ethics guidelines—and their implications for control or governance—in more

detail, we need to further develop the relationship between AI and societal norms and values.

As mentioned, how AI should be defined is a matter of considerable debate, which should be understood in relation to the technologies' expanding areas of use and the opportunities they offer, as well as their increasing everyday use. For example, Gasser and Almeida state that one reason for the difficulty in defining AI from a technical perspective is that it 'is not a single technology, but rather a set of techniques and sub-disciplines ranging from areas such as speech recognition and computer vision to attention and memory, to name just a few' (Gasser and Almeida 2017: 59). It is therefore apposite to highlight the relationship between, on the one hand, AI as a technological, mathematical and scientific concept and discipline and, on the other, the societal, normative and ethical values with which it interacts. This is particularly applicable to data-dependent machine learning in which the sample data used to train algorithms are based on human behaviour, norms and values. The latter half of this equation is studied by legal scholars, ethicists and other social scientists and humanists.

It is of particular interest from the perspective of AI's social applications, for example, in consumer markets and the public administration, to understand and evaluate machine-learning technologies based on societal values, norms and ethics. This gives rise to a multidisciplinary research requirement. Iyad Rahwan highlights exactly this interplay between technological development and societal values in terms of the need to keep 'society *in the loop*' in order to underpin the 'social contract' (see Fig. 4.1, Rahwan 2018). The term human-in-the-loop (HITL) is used in software development to describe a certain type of problem solving that includes people in the design of the solution. Spam filters and playlist recommendations, or indeed individual Facebook feeds, can be described as learning systems in which people's ongoing individual input plays a vital role in the problem-solving process itself: in deciding what is spam, which types of music the listener enjoys, which type of media and which friends the user finds most relevant. Rahwan raises this idea to a societal level.

Inspired by Rahwan's iterative concept, supplemented by the sociology of law and other social sciences perspectives, one can affirm that the need to constantly evaluate new self-learning technologies arises from at least three reasons:

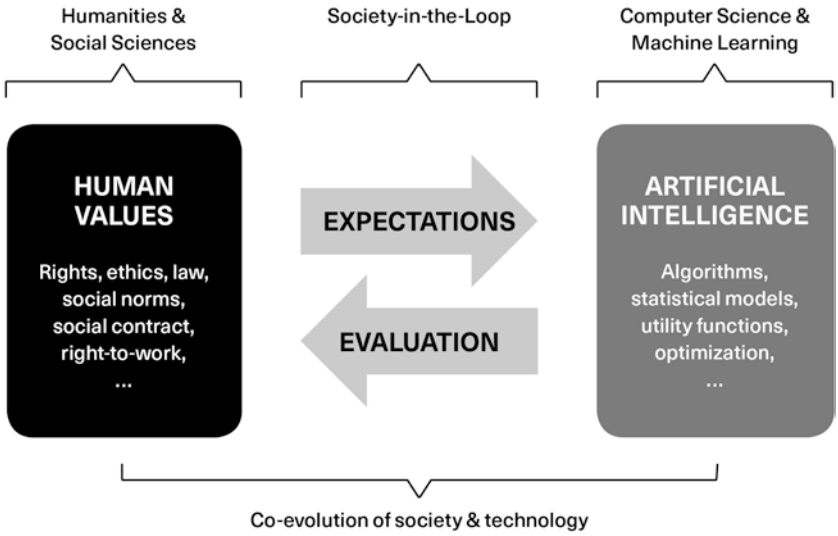

Fig. 4.1 Reproduced from Iyad Rahwan's 'Society-in-the-Loop: Programming the Algorithmic Social Contract' (2018)

1. *The multidirectional fairness concept*: To assess the fairness of the conclusions reached by self-learning technologies may be a computational, legal, cultural, normative or empirical venture, and the perspectives are not necessarily compatible (cf. Srivastava et al. 2019). Fairness may be viewed as an ethical or legal issue, studied across several disciplines. From one democratically influenced perspective, one might say that fairness is in part decided through a process of deliberation rather than an optimised expert process. That is, there may be two contrasting types of knowledge at play (cf. Lidskog 2008). From a legal point of view, however, society's quest for and construction of a well-balanced legal system is also a centuries-old undertaking. To normatively determine fairness is from this perspective not necessarily a calculable task but rather something that has taken the society a long time to establish structures to balance interests for. From yet another perspective, the normativity expressed in informal social norms can be seen as an empirical artefact—a Durkheimian social fact, or living law—to be studied and measured bottom-up.

2. *Human behaviour and social structures as training data*: On the one hand, there may be various reasons for skewed, biased or otherwise flawed datasets, albeit leading to a number of problematic outcomes, such as in the case with the study on geodiversity in ImageNet, mentioned earlier (Shankar et al. 2017). On the other, a normatively challenging question relating to accountability follows from the so-called mirroring effect (discussed in Larsson 2019), where it is not so much the data that are biased but the human societies' structures and human behaviour that are unfair, harmful or discriminatory. That is, shaping the underlying training data, the examples from which an AI-based system learns, leading to a reproduction or at worst amplification of racial, discriminatory, violent or otherwise harmful expressions. This guides the governance challenge back to the needs of transparency, as in auditability and detection, as well as questions of accountability.
3. *Misuse is always to be expected*. Powerful technologies will be used for a wide range of purposes, including criminal, fraudulent and repressive ones. For example, the threat scenarios outlined in a report on areas in which AI will be used for malicious ends included highly developed variations on cyberattacks such as automated hacking and the risk of remote takeovers of autonomous vehicles, which could thereby be utilised in physical attacks such as ramming crowds of people (Brundage et al. 2018). They also included the political and polarising use of bot networks to interfere in democratic elections, which has become a far too present destabilising feature of contemporary elections, it seems (cf. Bastos and Mercea 2019).

The misuses of GANs can also be mentioned in the third category. Initially focused on creating fake images of synthetic but realistic faces (Goodfellow et al. 2014), as the photorealism of the images has improved, GAN-based human image synthesis has been applied to creating deep fakes, sometimes for fraudulent purposes. The images have moved on from harmless applications such as creating realistic computer games and a speaking Mona Lisa, to being used for sinister purposes such as harassing women with manipulated naked photographs, a practice known as fake revenge porn (cf. Chesney and Citron 2019). In combination with the social phenomenon sometimes called the Streisand effect—when efforts to suppress online information result in wider dissemination and

longer-lasting exposure (cf. Jansen and Martin 2015)—the consequences of this type of harassment can be far-reaching, long-lasting and recurring.

Ethics Guidelines

A quick glance at the ongoing debate about AI and ethics at a global level confirms that there is currently a lively discussion in academic and policy-making circles. In particular, ethical guidelines have undergone remarkable development in the late 2010s. A study in 2019 collected at least 84 public-private initiatives that had produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI (Jobin et al. 2019). A total of 88 per cent of the guidelines had been released after 2016 and, for example, *transparency* was featured as a key concept in 73 of the 84 sources. Google and the Swedish telecom operator Telia are two examples of companies that have published ethical principles for their AI-based activities, while the relatively young AI Now Institute at New York University has become well known for its publications in the field. Others include the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore and the Select Committee on Artificial Intelligence of the UK House of Lords. Jobin et al. (2019) conclude that there is global *convergence* concerning at least five ethical principles: (i) transparency, (ii) justice and fairness, (iii) non-maleficence, (iv) responsibility and (v) privacy. Interestingly, at the same time, they find that there is *substantive divergence* regarding how these principles are interpreted: why they are deemed important; what issue, domain or stakeholders they pertain to; and how they should be implemented (Jobin et al. 2019).

As argued by Hagendorff (2020), the problem with ethical guidelines is that AI ethics—or ethics in general—lacks mechanisms to reinforce its own normative claims. According to Hagendorff, 'it is also a reason why ethics is so appealing to many AI companies and institutions. When companies or research institutes formulate their own ethical guidelines, regularly incorporate ethical considerations into their public relations work, or adopt ethically motivated "self-commitments", efforts to create a truly binding legal framework are continuously discouraged' (Hagendorff 2020: 100). He thereby places considerable emphasis on exactly this

avoidance of regulation as a principal objective of the AI industry's ethical guidelines. Mark Coeckelbergh expresses similar concerns in an article published in 2019 on the ethical issues and regulatory challenges of AI. Coeckelbergh, also a member of AI HLEG, states: 'There is a risk that ethics are used as a fig leaf that helps to ensure acceptability of the technology and economic gain but has no significant consequences for the development and use of the technologies' (Coeckelbergh 2019: 33). Even if he has a point—and it is undoubtedly the case that many companies flaunt their 'self-regulation' (in the form of relatively toothless internal policies) in order to avoid stricter external regulation—there may still be other reasons why ethics as an instrument of control has been so strongly emphasised in the field of AI development. The question is whether the specific nature of AI's rapid growth has played a significant role in this field, in particular requiring a softer approach while waiting for critical research to catch up and offer a stable foundation for potent regulation (cf. Larsson 2020). That said, there is also a question as to what the *juridification* of AI ethics might entail and which elements might be best suited—or ill-suited—for legislation (specifically on human oversight, see Koulu 2020).

The EU and Trustworthy AI

As mentioned in the Introduction, the newly appointed President of the European Commission, Ursula von der Leyen, expressed a willingness to introduce legislation in the area of the human and ethical implications of AI (2019). This is in line with that the EU adopted a strategy for AI in April 2018 and appointed the High-Level Expert Group on Artificial Intelligence, tasked with making policy and investment recommendations and offering guidance on ethical issues related to AI in Europe. In December 2018, the Commission presented its *Coordinated Plan on Artificial Intelligence 'Made in Europe'*, which was prepared in consultation with member states in order to promote the development and utilisation of AI in Europe. Among other things, the plan calls for all member states to have their own strategies in place by mid-2019. The High-Level Expert Group constellation did, however, not escape criticism; for example, Professor Yochai Benkler (2019) expressed a concern that representatives of industry are being given too much leeway in shaping AI regulation. Benkler drew parallels between the European Commission's expert group, Google's failed AI ethics council and Facebook's investments in a German research institute for AI ethics. The publication of the *Ethics Guidelines for*

Trustworthy AI even led to criticism from AI HLEG members themselves; in an interview, Thomas Metzinger, professor of theoretical philosophy at the University of Mainz, stated that the draft proposal prohibiting certain areas of use such as autonomous weapons systems or equivalents to the Chinese social credit system had been watered down by industry representatives and their allies.

While it remains to be seen what significance these sources will have for European AI development, the Commission has placed the expert group in a remarkably key position for influencing its future direction. AI HLEG is also the steering group for the European AI Alliance, a forum intended to engage various stakeholders in a broad and open dialogue on all aspects of AI development and its economic and social impact. In the Ethics Guidelines, the AI HLEG highlights three components of trustworthy AI that should be met throughout the AI system's entire life cycle. It should be:

(a) lawful, complying with all applicable laws and regulations;
(b) ethical, ensuring adherence to ethical principles and values; and
(c) robust, both from technical and social perspectives, since, even with good intentions, AI systems can cause unintentional harm.

The Ethics Guidelines deal with ethics (b) and robustness (c) but not legality (a), this despite the fact that, within the framework of ethics, the document inescapably deals with fundamentally legal issues in the form of accountability, transparency and, not least, data protection. As AI HLEG rightly says, much of the AI development and use in Europe falls within the scope of existing legislation. This applies to the Charter of Fundamental Rights, GDPR, the Product Liability Directive, Anti-Discrimination Directive, consumer protection legislation and so forth.

AI HLEG states seven main conditions for the implementation and realisation of trustworthy AI, all of which should be evaluated and managed throughout the AI system's life cycle (see Fig. 4.2).

1. Human agency and oversight: an AI system should be a source of a fair society by supporting human agency and fundamental rights, rather than reducing, limiting or undermining human independence.
2. Technical robustness and safety: trustworthy AI requires algorithms that are sufficiently secure, reliable and robust to deal with errors or inconsistencies in all phases of the AI system's work.

Fig. 4.2 From AI HLEG (2019b). Seven requirements for the realisation of trustworthy AI

3. Privacy and data governance: citizens should have full control over their personal data. This data must not be used to harm or disadvantage them.
4. Transparency: emphasises the AI system's traceability, explainability and communication.
5. Diversity, non-discrimination and fairness: AI systems should consider all degrees of people's talents, skills and needs, as well as guaranteeing user accessibility.

6. Societal and environmental well-being: AI systems should be utilised to reinforce positive social change and increase sustainability and environmental responsibility.
7. Accountability: mechanisms should be put in place to ensure responsibility and accountability for AI systems and the results of their processes, including the opportunity to review and report negative consequences.

With regard to the investment and policy recommendations published by AI HLEG (2019c), among other things, the group recommends a risk-based approach that is both proportional and effective in ensuring that AI is legal, ethical and robust in its adaption to fundamental rights. AI HLEG calls for a comprehensive mapping of relevant EU regulations and potential legal gaps to assess the extent to which various legal instruments continue to fulfil their purpose in an AI-driven world. The group underlines that new legal measures and governance mechanisms may need to be introduced to ensure adequate protection against negative effects and facilitate correct oversight and deployment.

An Ecosystem of Trust? Risk and Transparency

In February 2020, the European Commission published a White Paper on AI. While the White Paper on AI is too extensive to be thoroughly analysed here, a few key points can be made. First of all, the White Paper is heavily informed by both notions of excellence and trust, thus clearly influenced by AI HLEG. The White Paper particularly refers to lack of transparency as a key challenge for regulatory enforcement. The definitional struggles of AI are pointed to, in stating that 'the definition of AI will need to be sufficiently flexible to accommodate technical progress while being precise enough to provide the necessary legal certainty' in relation to new legal instruments (European Commission 2020: 16).

Levelled or Binary Risk-Approach?

Another point of particular debate regards the approach on governance of AI risks. The White Paper is taking a risk-based as well as sector-specific approach to ensure that 'the scope of the regulatory framework is targeted and provides legal certainty' (European Commission 2020: 17). In this, *high-risk applications* are distinguished from all other applications, but

pointing to healthcare, transport, energy and parts of the public sector as sectors where, given the characteristics of the activities typically undertaken, significant risks can be expected to occur (European Commission 2020: 17). In a cumulative fashion, thereby, the AI application in a high-risk sector should in addition be used in such a manner that significant risks are likely to arise, that is, pointing to both sector and intended use, for it to be addressed with a coming regulatory framework. The 'high-risk sector-requirement' alone has received critique. For example, it may miss the regulatory needs of potential detrimental effects on what would be called low-risk sector (cf. Dignum et al. 2020), which includes recommendations systems and targeted advertising, as these applications in large-scale risk leading to extreme effects in terms of polarisation, election influence or consumer manipulation. The German government, for example, has called for the proposed risk-classification system in the White Paper to be revised (Die Bundesregierung 2020).

The approach to risks with AI found in the White Paper can thereby be contrasted to regulatory discussions provided by the German Data Ethics Commission, which published a report on ethical guidelines at the end of October 2019. The German Data Ethics Commission makes a three-part distinction between algorithm-based decision-making, AI and data. Although the three are closely related and interlinked components, they still require individual focus. The work of the German Data Ethics Commission is in part governed by the same needs as AI HLEG, for example, with regard to human-centric design, privacy and self-determination, responsible data processing and the linking of digital strategies with sustainability goals. For 'algorithmic systems', the importance of transparency, explainability and clear divisions of responsibility is emphasised, and being in line with other guidelines.

Interestingly enough, the German Data Ethics Commission recommends adopting a risk-adapted regulatory approach to algorithmic systems, divided into five levels of criticality (2019:19–20). The principle underlying this approach means that the greater the potential for harm, the more stringent the requirements and the more far-reaching the intervention by means of regulatory instruments should be, that is, the higher are the requirements for transparency, inspection and evaluation. They are also open to a strict prohibition on the most high-risk applications. Risk assessment is one reasonable approach to sorting out requirements for regulation and methods of intervention. The German Data Ethics Commission proposes a labelling of algorithmic systems based on risk

assessments. Furthermore, it proposes the introduction of a national centre with specific expertise in algorithmic systems, tasked with assisting supervisory authorities in undertaking their mission. The German Data Ethics Commission also highlights the importance of explainable AI, implying a need to improve the comprehensibility of algorithmic systems in general and self-learning systems in particular.

The Multifaceted Transparency

Insights in the field of critical AI-related research have emerged relatively quickly, including that there are significant unintended negative consequences associated with socially integrated self-learning technologies. This also links to the difficulty in understanding and explaining certain outcomes obtained by what are sometimes referred to as black-box systems. As mentioned, these are all insights echoed by a majority of contemporary AI ethics guidelines (Jobin et al. 2019) as well as the work conducted by the High-Level Expert Group, emphasised in the European Commission's White Paper (2020) and the German Data Ethics Commission (2019). There are, however, several conflicts of interest linked to AI transparency (Larsson 2019; Larsson and Heintz 2020), and—as pointed out by Jobin et al. (2019)—what the concept itself would entail in relation to AI governance is diverging between different guidelines. In the literature, there are examples borrowing from the notion of environmental impact assessments. A report from the AI Now Institute focusing on the public sector summarises five key elements of public authority 'algorithmic impact assessment' (Reisman et al. 2018):

1. Agencies should conduct a self-assessment of existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias or other concerns across affected communities.
2. Agencies should develop meaningful external researcher review processes to discover, measure or track impacts over time.
3. Agencies should provide notice to the public disclosing their definition of 'automated decision system', existing and proposed systems and any related self-assessments and researcher review processes before the system has been acquired.
4. Agencies should solicit public comments (consultation) to clarify concerns and answer outstanding questions (dialogue).

5. Governments should provide enhanced due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased or otherwise harmful system uses that agencies have failed to mitigate or correct.

These recommendations for an 'algorithmic impact assessment' echo parts of the German Data Ethics Commission and emphasise participatory aspects of consultation known from the impact assessment literature.

From a governance perspective, there is a considerable difference between the needs and regulation of public administration compared to the scalability and multi-jurisdictional nature of global mega-platforms. Coeckelbergh, mentioned earlier, points out the difficulty in implementing ethical AI in relation to mega-platforms (cf. Lundqvist's chapter in this book), stating that 'it is hard to see how responsible innovation can really be implemented when there is a concentration of power in the hands of a relatively limited number of powerful actors, including a small number of large corporations: it seems that a handful of companies decide the future of AI' (Coeckelbergh 2019: 33). To understand how individuals come into day-to-day contact and interact with applied artificial intelligence—such as facial recognition, targeted marketing based on automated analytical inferences and semi-automated content moderation—and what the conditions are for regulation and implementing ethical guidelines in these contexts, we need to look at the tensions and power structures at work, not least vis-à-vis the large-scale digital platforms (cf. van Dijck et al. 2018).

Summing up: From Principles to Effective Implementation

In the preceding sections, I have argued that the difficulty in defining AI is one of the regulatory challenges concomitant to the application and development of artificial intelligence. My contention is primarily based on the need to view today's machine-learning-based, data-dependent AI in relation to societal structures and human values. One reason is that, for many types of applications, it is human expression—facial recognition, GPS data, social media behaviour and so forth—that constitutes the vast quantities of training data on which the precision of AI applications depends. This implies that while the regulatory challenge lies in the power and potential agency of AI methods, it also lies in the fact that these

methods reproduce society's imbalances. There is a risk that the data-dependency of current machine-learning technologies, in combination with the technology-driven market complexities that hamper explainability, transparency and supervisory oversight, will result in society's biases not only being reproduced but also amplified, while at the same time the harmful effects will be difficult to detect. It is also true that the power of these technologies, which provides precision in applications such as image analysis, behavioural prediction or the ability to generate synthetic data, has increased strongly in a very brief period of time. This in itself creates a regulatory challenge in terms of a 'lag', given that the consequences of this rapid increase take time to understand and evaluate (cf. Larsson 2020). In turn, legislative processes demand reflection and deliberation to achieve the desirable and reasonable balancing between the various social interests to which today's AI development relates in order to counteract challenging aspects such as lack of individual self-determination, information asymmetries, imbalances of power and risks for discrimination and manipulation. This temporal aspect, in combination with the balancing of multiple interests, is probably one significant explanation for why, at the beginning of the 2020s, governance in this field is largely characterised by ethical guidelines. In the light of the themes addressed in this chapter, there are at least three key questions to focus on going forward (explained in the following text).

From Principle to Procedure

The flora of ethical guidelines for AI development and use is rich in principled positions but poor in procedural arrangements. It is reasonable to assume that this is a matter of maturity, with principled consideration as the necessary first stage; however, the subsequent procedural stage is necessary both to strengthen the ability to implement those principled positions and to ensure trustworthy AI that citizens, public authorities, consumers and businesses trust to use, rely on and invest in. If one understands the growth of ethical guidelines as an expression of the rapid development of AI methods, then this can be seen as the expected second procedural stage. If, on the other hand, one views ethical guidelines as the result of corporate recalcitrance in the face of regulation and a soft and intentionally toothless alternative to legislation, then the procedural stage will meet with resistance. In the long-term, this procedural stage also implies an appeal for supervisory authorities to develop their methods as

part of the practical implementation of existing regulation (cf. Larsson 2018). That which AI HLEG expresses as a need for 'auditability' can, in part, be transferred to supervisory authorities with responsibility for ensuring market function, which is highly necessary given the monopolistic tendencies of global platforms and the complex ecosystem of data-driven markets. The AI-driven elements of markets that can control individualised ad distribution, pricing and the like need to be auditable, something that demands methodological development. In this vein, for example, the German Data Ethics Commission proposes a central expert group tasked with supporting supervisory authorities in this task.

The Need for a Multidisciplinary Approach

A core argument in this chapter regards the necessity of a multidisciplinary approach in the studies and understandings of issues of AI governance. It is reasonable to assume that meeting many of the knowledge needs will demand collaboration between mathematics-rooted computer science disciplines, with their deep insight into the construction and workings of AI systems, and disciplines in social sciences and the humanities, which are in a position to theorise and increase the understanding of AI's interactions with human cultures, norms, values and attitudes or its implications for power, markets, states and regulations. The interaction that characterises a *society-in-the-loop*, as discussed earlier (Rahwan 2018), in which human expression and social structures constitute AI's training data, gives rise to normative questions about AI systems, where balance of interests and human values are emphasised. Somewhat counter-intuitively, many of the challenges posed by applied AI systems regard the interplay between an adaptive technology and human behaviour and societal structures. The human-centric approach often cited in ethical guidelines for AI may contain a somewhat idealised conception of what humanity means in terms of values and social structures. Conversely, from an empirical behavioural approach, one might say that it is often the application of (bad) human behaviour and skewed social structures that leads to automated failure. For a data-dependent AI learning from vast quantities of training data, this quite simply means that lessons will be taken not just from good and well-balanced examples but also from humanity's less noble side: its racism, xenophobia, gender inequalities and informal but widespread structural injustice. The challenge here is therefore the *normative sorting* of underlying data or, alternatively, to weight self-learning technologies' automation

and scalability to compensate for their reproductive and amplifying tendencies, so that their outcomes are more balanced than the underlying data, or produce more fair outcomes than the human structures at present offer. That is, to normatively aspire for more, and therefore unavoidably become entangled with all the value-based messiness of normative plurality.

Legal Adaptability in Times of Change

Most certainly, more sectoral guidelines, as well as clarified processes for implementation and supervision over applied AI, can be expected. However, there is an adaptability to core legal concepts that can be expected to reveal just how flexible they are. History teaches us that regulation in times of rapid technology-driven social change may pose significant challenges. That said, legal scholars such as Karl Renner, who analysed property regulation during Western Europe's industrialisation, teach us that the law itself can be incredibly dynamic. Given the pronouncement by European Commission President Ursula von der Leyen, followed by the White Paper, a regulatory process is on its way, even if its exact contours are yet to be fixed. The interpretation of existing national and European legislation in the light of the functionality, possibilities and challenges of AI systems is currently at hand, also motivated by member states' drafting of national AI strategies. This may be a challenging task, partly because the rapid pace of technology-driven change is difficult to get to grips with given the protracted nature of traditional regulation and partly because we are still building up our knowledge of the fundamental consequences of contemporary AI in and for society.

The issue of AI has taken a value-based, ethics-centred turn within Europe. It is a question of how we view those qualities of AI that I find most laudable; the precision of self-learning and autonomous technologies must be assessed in their interaction with societal values—*in-the-loop*. This is a normative definition, with a bearing on future lines of development: a good AI is an AI rooted in human society and one we can trust.

References

Abbott, K. W., & Snidal, D. (2000). Hard and Soft Law in International Governance. *International Organization, 54*(3), 421–456.

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review, 37*(1), 38–54.

Benkler, Y. (2019). Don't Let Industry Write the Rules for AI. *Nature, 569*(7754), 161–162.

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. & Anderson, H. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:*1802.07228.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency, PMLR, 81*, 77–91.

Chesney, B., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review, 107*, 1753.

Coeckelbergh, M. (2019). Artificial Intelligence: Some Ethical Issues and Regulatory Challenges. *Technology and Regulation*, 31–34. https://doi.org/10.26116/techreg.2019.003.

Cotterrell, R. (1992). *The Sociology of Law: An Introduction*. Oxford: Oxford University Press.

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies, 2015*(1), 92–112.

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology, 31*(4), 525–541.

de Vries, K. (2020). You Never Fake Alone. Creative AI in Action. *Information, Communication & Society*, *23*(14), 2110–2127.

Die Bundesregierung. (2020, June 29). *Stellungnahme der Bundesregierung der Bundesrepublik Deutschland zum Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen*. COM (2020) 65 Final.

Dignum, V., Muller, C. & Theodorou, A. (2020). First Analysis of the EU Whitepaper on AI. ALLAI. https://allai.nl/first-analysis-of-the-eu-whitepaper-on-ai/

Ellickson, R. C. (1994). *Order without Law*. Cambridge, MA: Harvard University Press.

European Commission. (2018, April 25). *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 on Artificial Intelligence for Europe*. COM (2018) 237 final.

European Commission. (2020, February 19). *White Paper on Artificial Intelligence: Public Consultation Towards a European Approach for Excellence and Trust*. COM (2020) 65 final.

Fuller, L. L. (1975). Law as an Instrument of Social Control and Law as a Facilitation of Human Interaction. *BYU Law Review, 1*, 89–98.

Gasser, U., & Almeida, V. A. (2017). A Layered Model for AI Governance. *IEEE Internet Computing, 21*(6), 58–62.

German Data Ethics Commission. (2019). *Opinion of the Data Ethics Commission*. Retrieved 20 September 2020 from https://datenethikkommission.de/wp-content/uploads/DEK_Gutachten_engl_bf_200121.pdf

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines, 30*, 99–120.

HLEG. (2019a). *A Definition of AI: Main Capabilities and Disciplines: Definition Developed for the Purpose of the AI HLEG's Deliverables*. Brussels: European Commission.

HLEG. The (2019b). *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission.

HLEG. (2019c). *Policy and Investment Recommendations for Trustworthy Artificial Intelligence*. Brussels: European Commission.

Jansen, S. C., & Martin, B. (2015). The Streisand Effect and Censorship Backfire. *International Journal of Communication, 9*, 656–671.

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Koulu, R. (2020). Human Control over Automation: EU Policy and AI Ethics. *European Journal of Legal Studies, 12*, 9–46.

Larsson, S. (2017). *Conceptions in the Code. How Metaphors Explain Legal Challenges in Digital Times*. New York: Oxford University Press.

Larsson, S. (2018). Algorithmic Governance and the Need for Consumer Empowerment in Data-Driven Markets. *Internet Policy Review, 7*(2), 1–12.

Larsson, S. (2019). The Socio-Legal Relevance of Artificial Intelligence. 'Law in an Algorithmic World'. Special Issue of *Droit et Société, 103*(3), 573–593.

Larsson, S. (2020). On the Governance of Artificial Intelligence Through Ethics Guidelines. *Asian Journal of Law and Society, 1*, 1–15.

Larsson, S., & Heintz, F. (2020). Transparency in Artificial Intelligence. *Internet Policy Review, 9*(2), 1–16.

Legg, S., & Hutter, M. (2007) A Collection of Definitions of Intelligence. In B. Goertzel & P. Wang (Eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* (pp. 17–24), Proceedings of the AGI Workshop 2006 (Vol. 157), IOS Press.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-Making Processes. *Philosophy & Technology, 31*, 611–627.

Lidskog, R. (2008). Scientised Citizens and Democratised Science. Re-assessing the Expert-Lay Divide. *Journal of Risk Research, 11*(1–2), 69–86.

Mandel, G. N. (2009). Regulating Emerging Technologies. *Law, Innovation and Technology, 1*(1), 75–92.

Pasquale, F. (2015). *The Black Box Society. The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Pro Publica. (2016, May 23). *Machine Bias.* Retrieved September 22, 2020, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Rahwan, I. (2018). Society-in-the-Loop: Programming the Algorithmic Social Contract. *Ethics and Information Technology, 20*(1), 5–14.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. *AI Now Institute*, 1–22.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development, 3*(3), 210–229.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv preprint arXiv:*1711.08536.

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2459–2468).

Stone, P., et al. (2016). *Artificial Intelligence and Life in 2030, Report of the* (pp. 2015–2016). Stanford University: Study Panel. Stanford.

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The Platform Society: Public Values in a Connective World*. New York: Oxford University Press.

von der Leyen, Ursula. (2019). *A Union That Strives for More. My Agenda for Europe. Political Guidelines for the Next European Commission 2019–2024.* Retrieved 20 September 2020 from https://op.europa.eu/en/publication-detail/-/publication/43a17056-ebf1-11e9-9c4e-01aa75ed71a1

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive Inequity in Object Detection. *arXiv preprint arXiv:*1902.11097.