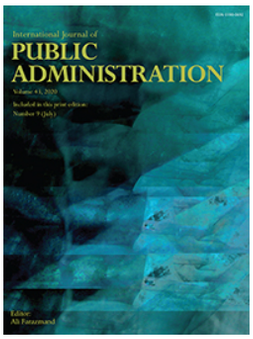

International Journal of Public Administration

ISSN: 0190-0692 (Print) 1532-4265 (Online) Journal homepage: www.tandfonline.com/journals/lpad20

The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration

Bernd W. Wirtz, Jan C. Weyerer & Benjamin J. Sturm

To cite this article: Bernd W. Wirtz, Jan C. Weyerer & Benjamin J. Sturm (2020) The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration, International Journal of Public Administration, 43:9, 818-829, DOI: 10.1080/01900692.2020.1749851

To link to this article: https://doi.org/10.1080/01900692.2020.1749851

INTERNATIONAL JOURNAL OF PUBLIC ADMINISTRATION
2020, VOL. 43, NO. 9, 818–829
https://doi.org/10.1080/01900692.2020.1749851

Routledge
Taylor & Francis Group

The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration

Bernd W. Wirtz, Jan C. Weyerer, and Benjamin J. Sturm

Chair for Information and Communication Management, German University of Administrative Sciences Speyer, Speyer, Germany

ABSTRACT

As government and public administration lag behind the rapid development of AI in their efforts to provide adequate governance, they need respective concepts to keep pace with this dynamic progress. The literature provides few answers to the question of how government and public administration should respond to the great challenges associated with AI and use regulation to prevent harm. This study analyzes AI challenges and former AI regulation approaches. Based on this analysis and regulation theory, an integrated AI governance framework is developed that compiles key aspects of AI governance and provides a guide for the regulatory process of AI and its application. The article concludes with theoretical implications and recommendations for public officers.

KEYWORDS

Artificial intelligence; governance; regulation; framework; public administration; regulation theory; AI challenges

Introduction

"A robot may not injure a human being or, through inaction, allow a human being to come to harm." (Asimov, 1950, p. 26). With this famous quote, Isaac Asimov was among the first to present a set of rules for living in a society with intelligent machines. Today, this fiction from the past becomes more and more reality, as scientists and engineers are working hard to create artificial intelligence (AI), which is "the capability of a computer system to show human-like intelligent behavior characterized by certain core competencies, including perception, understanding, action, and learning" (Wirtz et al., 2019, p. 599).

AI provides great opportunities for public administration, including the automation of workflow processes, faster information processing, improved service quality or increased working efficiency (Thierer et al., 2017; Zheng et al., 2018). Due to these potential benefits, government and public administration increasingly acknowledge the significance of AI for economic and social advancement by applying AI to their administration and public infrastructures, as well as by supporting AI research.

Despite these great opportunities, still many challenges and risks are associated with implementing AI in public administration, constituting a darker side of AI. These challenges vary from issues of data privacy and security, to workforce replacement and ethical problems like the agency and fairness of AI (Boyd & Wilson, 2017; Wang & Siau, 2018). Prominent opinion leaders have recently intensified their efforts to raise awareness for these challenges and threats. Experts like Elon Musk and Stephen Hawking have brought up the question, whether AI will always be beneficial or rather become a threat to humanity (Cuthbertson, 2018). While these questions have also recently become subject to AI research, only few studies have elaborated on them. Accordingly, there is little knowledge about the challenges of AI associated with the public sector and no consensus about how to deal with them in the future (Veale et al., 2018; Wang & Siau, 2018). However, Scherer (2016) highlights the need for a legal system assessing benefits and risks to find a way to regulate AI and respective research without interfering with its advancement. Boyd and Wilson (2017) emphasize the need for local and international policies to reduce social and personal risks caused by AI. Thus, there are efforts to find global solutions to these challenges, but many governments and researchers struggle in formulating a long-term perspective on how to regulate and interact with the AI market in both the private and public sector (Cath et al., 2017; Scherer, 2016).

Against this background, this article first outlines the current state of AI governance before giving an overview of AI challenges and risks for public administration as well as previous AI governance or regulation frameworks. Based on this analysis and regulation theory, an integrated AI governance framework is developed that organizes the key aspects of AI governance and regulation, showing the complex interactions of AI challenges and their regulation in the context of public administration. The last section discusses theoretical

CONTACT Bernd W. Wirtz ls-wirtz@uni-speyer.de Chair for Information and Communication Management, German University of Administrative Sciences Speyer, Freiherr-vom-Stein-Str. 2, Speyer 67346, Germany.

and practical implications of the findings, revealing opportunities for future research and providing recommendations for public officers.

Current state of AI governance

As the number of AI applications is growing and the technology increasingly permeates everyday life, the question arises of how government and public administration should deal with the potential risks and challenges involved, which is currently heavily discussed in the media and scientific literature (Boyd & Wilson, 2017; Smith, 2018). As governance greatly affects the AI industry, the development of AI and its impact on society, Thierer et al. (2017) propose two different approaches for governing AI. The first method uses restrictions, bans and prohibitions to limit research on AI as well as its application in any public or private environment to prevent potential harm from autonomous machines or AI algorithms. While this precautionary principle could severely impede technological progress and deprive society of possible benefits, governance could take, on the other side, a reactive approach of unrestricted innovation, preventing and regulating risks only when they occur in reality.

Some governments are already taking action, providing financial support to fund new innovation and technology and planning strategies on how to interact, regulate and govern AI in the future (Ansip, 2017). Likewise, private organizations such as the Institute of Electrical and Electronics Engineers (IEEE, 2017), the Allianz Group (AGCS, 2018) or Microsoft (Smith, 2018) discuss the impact of AI on society and possible ways of governance.

In the literature, the topic of AI governance and regulation is widely unexplored. Scherer (2016) gives a detailed analysis of AI challenges and the theoretical role of the government in legal questions, proposing a legal system that involves the invention of an AI development act consisting of a predefined legal rule system and a newly created agency to control other organizations and enforce these rules. While there are many frameworks in the literature about the governance of IT or organizations in general, only few focus on AI. Most of these models concern either technical or structural aspects of AI (Sirosh, 2017), focus on organizational implementation (Bataller & Harris, 2016) or address the process of implementing AI into a public organization (Zheng et al., 2018). Only two frameworks address the governance or regulation of AI risks and challenges (Gasser & Almeida, 2017; Rahwan, 2018). The model proposed by Gasser and Almeida (2017) consists of three hierarchical layers. The first layer covers technical aspects of AI technology, algorithms and data structures. This is the core of their model, as AI systems are based on algorithms and are processing data to make decisions or take actions. The authors further propose principles of responsibility and explainability to secure fairness and non-discriminatory actions at this early stage of information processing. The second layer addresses ethical criteria and principles to be considered and used to design specific ethics for the use of AI. The third layer covers social and legal issues and demands a regulatory framework, defining challenges and responsibilities for AI to generate norms and appropriate regulation and legislation in the long term. Although, the model gives a decent overview of the different domains that are in need of regulation, it fails to provide details on how the different layers interact, how to put the regulatory process into practice and who should be responsible for the proposed regulation or governance. The model also solely relies on the challenging aspects of AI without a concrete theoretical foundation.

Another approach by Rahwan (2018) proposes a societal contract to evaluate the behavior of AI technology via a 'society-in-the-loop', an extension of the human-in-the-loop approach. Human-in-the-loop describes an approach to AI technology, where a human operator is always supervising and managing the outputs and actions of the AI system. The purpose of the AI system is to make recommendations, but the final decision on how to act is always based on information that comes from the human operator who thus controls the actions of AI to achieve the common goals of the stakeholders involved. According to Rahwan (2018), the human-in-the-loop approach is not sufficient for regulating the future use of AI, as AI will be applied to broader areas with implications for the whole society and requires a regulation approach that represents all stakeholders within society, even though their interests might interfere with each other. In his society-in-the-loop model, the society as a whole first has to resolve certain tradeoffs between different human values like privacy and safety and has to ensure that the benefits and costs of AI technology are reasonably distributed among the different stakeholders. Government and industry are meant to work together to provide regulations and standards and represent the goals and expectations of society based on human values, ethics and social norms. The results are then implemented into an AI algorithm and evaluated towards the defined goals. In this case, all parts of society collaborate to govern or regulate the goals and behaviors of AI technology. As can be seen, the model of Rahwan (2018) considers a broader view of AI in society. It accounts for different stakeholders and conflicting interests that are to be resolved within a collaborative effort to increase benefits to society. Although some limitations like the ability to quantify social

values are discussed, it lacks details on how to design and implement AI regulation. While the model also provides some information about the actors in the societal loop, it neglects the responsibilities of government.

Both models discuss the reasons for regulation as well as the challenges and risks of AI only superficially and provide little information and theoretical reasoning for the actual necessity of regulation. However, to develop a thorough integrated AI governance framework requires an understanding and examination of these risks and challenges on a deeper level to be able to target specific aspects without interfering with or slowing down technical innovations and progress with incomplete or too general policies (Thierer et al., 2017).

Overview of AI challenges in the literature

To present a systematic overview of the current state of literature on AI challenges and risks as well as to define areas in need of regulation, three main areas of public AI challenges are emphasized based on the recent AI challenges approach of Wirtz et al. (2019) (see Figure 1).

AI law and regulation

This area strongly focuses on the control of AI by means of mechanisms like laws, standards or norms that are already established for different technological applications. Here, there are some challenges special to AI that need to be addressed in the near future, including the governance of autonomous intelligence systems, responsibility and accountability for algorithms as well as privacy and data security.

*Governance of autonomous intelligence system*addresses the question of how to control autonomous systems in general. Since nowadays it is very difficult to conceive automated decisions based on AI, the latter is often referred to as a 'black box' (Bleicher, 2017). This black box may take unforeseeable actions and cause harm to humanity. For instance, if an autonomous AI weapon system learned that it is necessary to prevent all threats to obtain security, it might also attack civilians or even children classified as armed by the opaque algorithm (Heyns, 2014). Situations can get even worse when the AI becomes autonomous enough to pursue its own goals, even if this means harm to individuals or humanity (Lin et al., 2008). Examples like this give rise to the questions of transparency and accountability for AI systems.

The challenge of *responsibility and accountability* is an important concept for the process of governance and regulation. It addresses the question of who is to be held legally responsible for the actions and decisions of AI algorithms. Although humans operate AI systems, questions of legal responsibility and liability arise. Due to the self-learning ability of AI algorithms, the operators or developers cannot predict all actions

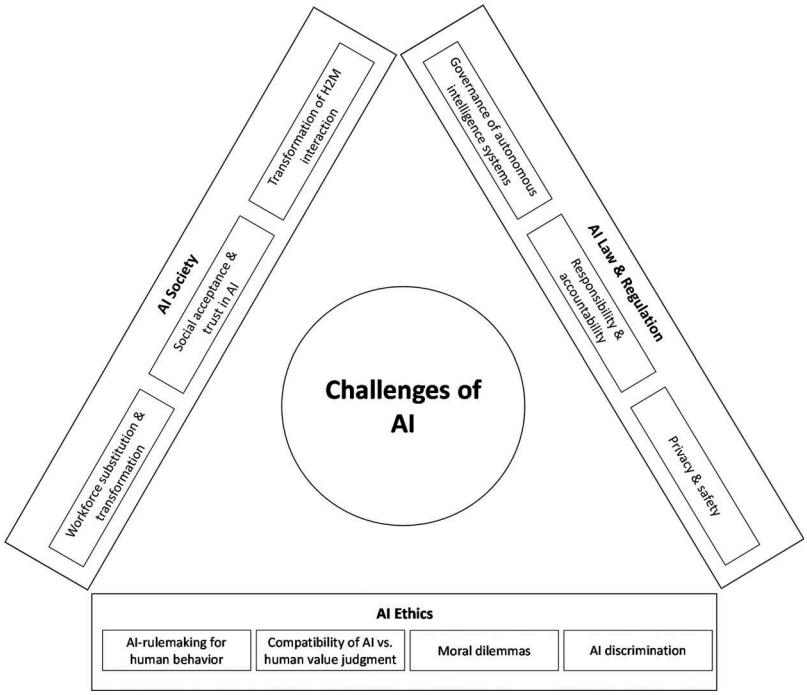

Figure 1. The three main areas of public AI challenges.

and results. Therefore, a careful assessment of the actors and a regulation for transparent and explainable AI systems is necessary (Helbing et al., 2017; Wachter et al., 2017).

Privacy and safety deals with the challenge of protecting the human right for privacy and the necessary steps to secure individual data from unauthorized external access. Many organizations employ AI technology to gather data without any notice or consent from affected citizens (Coles, 2018). For instance, when searching for a fast way to get home from work, a navigation system has to access the current location of the user or the government uses AI services to monitor public spaces to prevent criminal activities (Power, 2016). Without informed consent from the affected individuals, these AI applications and services endanger their privacy.

AI society

AI already shapes many areas of daily life and thus has a strong impact on society and everyday social life. For instance, transportation, education, public safety and surveillance are areas where citizens encounter AI technology (Stone et al., 2016; Thierer et al., 2017). Many are concerned with the subliminal automation of more and more jobs and some people even fear the complete dependence on AI or perceive it as an existential threat to humanity (McGinnis, 2010; Scherer, 2016).

Workforce transformation and substitution is an important topic for government, industry and society as a whole (Wirtz et al., 2019). AI can reduce tedious and repetitive work and is able to save time for the user for more creative or difficult to automate tasks. Furthermore, the accuracy in data analysis has improved to a point, where it is better than the human ability (Esteva et al., 2017). Frey and Osborne (2017) analyzed over 700 different jobs regarding their potential for replacement and automation, finding that 47 percent of the analyzed jobs are at risk of being completely substituted by robots or algorithms. This substitution of workforce can have grave impacts on unemployment and the social status of members of society (Stone et al., 2016).

Social acceptance and trust in AI is highly interconnected with the other challenges mentioned. Acceptance and trust result from the extent to which an individual's subjective expectation corresponds to the real effect of AI on the individual's life. In the case of transparent and explainable AI, acceptance may be high but if an individual encounters harmful AI behavior like discrimination, acceptance for AI will eventually decline (COMEST, 2017). To reduce negative attitudes towards AI, government and industry can influence social acceptance with good governance and standards to enforce its beneficial use (Scherer, 2016).

Human interaction with machines is a big challenge to society because it is already changing human behavior. Meanwhile, it has become normal to use AI on an everyday basis, for example, googling for information, using navigation systems and buying goods via speaking to an AI assistant like Alexa or Siri (Mills, 2018; Thierer et al., 2017). While these changes greatly contribute to the acceptance of AI systems, this development leads to a problem of blurred borders between humans and machines, where it may become impossible to distinguish between them. Advances like Google Duplex were highly criticized for being too realistic and human without disclosing their identity as AI systems (Bergen, 2018).

AI ethics

Ethical challenges are widely discussed in the literature and are at the heart of the debate on how to govern and regulate AI technology in the future (Bostrom & Yudkowsky, 2014; IEEE, 2017; Wirtz et al., 2019). Lin et al. (2008, p. 25) formulate the problem as follows: "there is no clear task specification for general moral behavior, nor is there a single answer to the question of whose morality or what morality should be implemented in AI". Ethical behavior mostly depends on an underlying value system. When AI systems interact in a public environment and influence citizens, they are expected to respect ethical and social norms and to take responsibility of their actions (IEEE, 2017; Lin et al., 2008).

AI rulemaking for humans can be the result of the decision process of an AI system when the information computed is used to restrict or direct human behavior. The decision process of AI is rational and depends on the baseline programming. Without the access to emotions or a consciousness, decisions of an AI algorithm might be good to reach a certain specified goal, but might have unintended consequences for the humans involved (Banerjee et al., 2017).

AI discrimination is a challenge raised by many researchers and governments and refers to the prevention of bias and injustice caused by the actions of AI systems (Bostrom & Yudkowsky, 2014; Weyerer & Langer, 2019). If the dataset used to train an algorithm does not reflect the real world accurately, the AI could learn false associations or prejudices and will carry those into its future data processing. If an AI algorithm is used to compute information relevant to human

decisions, such as hiring or applying for a loan or mortgage, biased data can lead to discrimination against parts of the society (Weyerer & Langer, 2019).

Moral dilemmas can occur in situations where an AI system has to choose between two possible actions that are both conflicting with moral or ethical values. Rule systems can be implemented into the AI program, but it cannot be ensured that these rules are not altered by the learning processes, unless AI systems are programed with a "slave morality" (Lin et al., 2008, p. 32), obeying rules at all cost, which in turn may also have negative effects and hinder the autonomy of the AI system.

Compatibility of machine and human value judgment refers to the challenge whether human values can be globally implemented into learning AI systems without the risk of developing an own or even divergent value system to govern their behavior and possibly become harmful to humans. To prevent this potential threat, unchangeable rules would have to be implemented, leading to less autonomy of the AI and the above-mentioned problem of slave morality (Lin et al., 2008). The deliberations and findings from previous literature show that implementing and applying AI involves great challenges and risks for the public, calling for reasonable and beneficial governance or regulation concepts.

Regulation theory as a basis for an AI governance framework

The main purpose of governance is to enable institutions, society and other stakeholders to work together and fulfill policy goals in a dynamic and changing environment without grave interruptions or damage to society (Asaduzzaman & Virtanen, 2016). From a political-economic point of view, the described negative interruption and adverse effects of AI may be viewed as a market failure (Rahwan, 2018). Market failure is a core concept of regulation theory and refers to a state in a free market, where resources are not efficiently allocated and not all costs and opportunities are considered for all stakeholders in the market (De Geest, 2017). Market failure often results in harmful situations and instability for society, leading civil stakeholders to ask for regulation (Stemler, 2016). One important factor contributing to market failures are externalities or external effects. Externalities are benefits or costs that occur outside of the initial market without any form of compensation or recognition (Delucci, 2000). AI applications, technologies and services can cause such negative external effects and thus contribute to market failures (Rahwan, 2018). A common example in the AI context refers to technological unemployment through AI-based industry robots. Therefore, the AI and its corresponding applications and services can be regarded as the object of regulation.

According to traditional regulation theory, the government is responsible for regulation to avoid market failures and prevent harm to members of society. Since the infancy of public AI makes it nearly impossible to explain and enact regulation based on past occurrence or historical judgments, many governments seek to base regulation on inherently normative elements such as human values and ethics to guide regulation attempts (Cath et al., 2017). Normative regulation theory, which focuses on an ideal form of regulation and how to influence the market to become most efficient (Den Hertog, 2012), therefore appears to be particularly suited to approach AI regulation and explain its basic modus operandi.

From an interest-based regulation-theoretical perspective, governmental regulation is justified with the protection of the public interest against other forms of interest. Regulation is used to enable optimal allocation of resources and a stable market for all participants or stakeholders involved (Baldwin et al., 2012), and thus to prevent the occurrence of market failures. The deciding regulator has to manage the different stakeholder expectations to maximize the own benefit. In other words, "regulators attempt to strike a balance that society is comfortable with through a constant learning process" (Rahwan, 2018, p. 10). In this way, an equilibrium between public and private interest is realized in the regulatory process (Baldwin et al., 2012), which thus represents an essential component of regulation theory.

To enhance the regulatory process, Boyd and Wilson (2017) argue that the actors in this process should collaborate to represent the knowledge and expertise of all interest groups. Accordingly, the different actors in this policy-making process and their collaboration may play a significant role in the context of regulation.

As regulation can have both positive and negative effects, the actors need to consider carefully the different methods of regulation and implementation. The regulatory process therefore has to account not only for the challenges posed by AI, but also for possible increases in efficiency, reduction of work load and other beneficial effects (Stone et al., 2016). Thierer et al. (2017, p. 54) argue that "[t]he benefits of AI technologies are simply too great for us to allow them to be extinguished by poorly considered policy". This stresses the importance of a regulatory process able to assess benefits, risks and challenges to generate

a beneficial AI policy as an outcome. Based on the above-mentioned regulation-theoretical deliberations, five core elements may be deduced for the development of an AI governance framework. These elements include the reason for regulation, the object of regulation, the regulatory process itself, the actors involved in the process and the outcome of the regulatory process.

An integrated AI governance framework

The conceptual framework combines insights from AI challenges and governance literature with the implications of regulation theory. The regulation-theoretical deliberations and core elements identified, determine the basic structure of the framework that consists of five layers. (1) As AI technology, services and applications are able to cause market failures, they represent the objective of regulation (AI technology, services and applications layer). (2) Market failure manifests itself through an external effect of the AI technology and the associated challenges posed to society (AI challenges layer). (3) To counter possible negative effects, a regulatory process is needed to assess costs and benefits as well as to evaluate the outcomes with and without regulation (AI regulation process layer). (4) At the end of the process, policies, laws and other means of regulation are implemented to prevent or adjust the aspects leading to market failure (AI policy layer). (5) Given the great impact of regulation on society and its potentially negative effects, the affected stakeholders and representatives of public and private interest groups should support the entire regulatory process (Collaborative AI governance layer). Figure 2 depicts the integrated AI governance framework with its individual layers.

AI applications/services and technology layer

This layer conceptualizes the realm of AI, how it gathers and processes data to reach its result. To think about the processes in an AI system, Bataller and Harris (2016) proposed the terms of sensing, comprehending and acting on given data. Sensing AI describes all its efforts and parts of data acquisition. This resembles a process of perception, in which information from the application environment of the AI are gathered. This perception mostly occurs via the search or integration of an already acquired dataset or database, but also via external sensors, like cameras, tactile sensors or microphones.

After acquiring the data necessary for a task, the data needs to be further processed for the purpose of comprehension. Therefore, the AI algorithm has to structure new information and integrate it with all information gained before to build up a representation of virtual knowledge. This knowledge is analyzed for possible patterns afterwards to derive conclusions about the content of new and old pieces of information. In the last step of acting, the gained information from the second step is used to perform a certain action.

Depending on the purpose of the AI system in question, this action could refer to the adjustment of the own learning algorithm, to the report of insights drawn from the data (e.g., financial report) or the decision to act in a certain way, like steering an automated car to the left side to evade a cyclist on the street. Since it is possible to encounter challenges of AI at each of these steps, it is important to distinguish between these different stages of processing to implement the right regulation method.

AI challenges layer

The reason for regulation are the challenges elicited by the AI technology and services described above. The AI challenges layer is divided into three parts: AI society, AI ethics and AI law and regulation. Even though these areas are interconnected, the framework treats them as distinct parts, as each of them requires a different analysis of the underlying problems and different forms of regulation. For example, laws can be implemented to secure a certain degree of data privacy, but it is very difficult to use a law to counteract all possible ethical or moral dilemmas that can occur. Ethical questions are mostly answered by acknowledging certain principles or norms that can be transformed or translated into enforceable laws in the aftermath (IEEE, 2017).

AI regulation process layer

As mentioned earlier, many models of AI regulation give broad overviews but fail to provide details on how to implement regulation in society. Similar to the approach of König et al. (2010), the regulatory process proposed in this framework comprises the concepts of framing, risk and benefit assessment, risk evaluation and the risk management.

The first step in the regulation process is *framing*. In this planning stage, stakeholders interested in the regulation of certain challenges come together to formulate a common understanding of the problem, defining the objective of the regulatory action they want to enact. To reach this common problem definition, the specific challenges have to be assessed with regard to risks, benefits and costs for stakeholders and the society

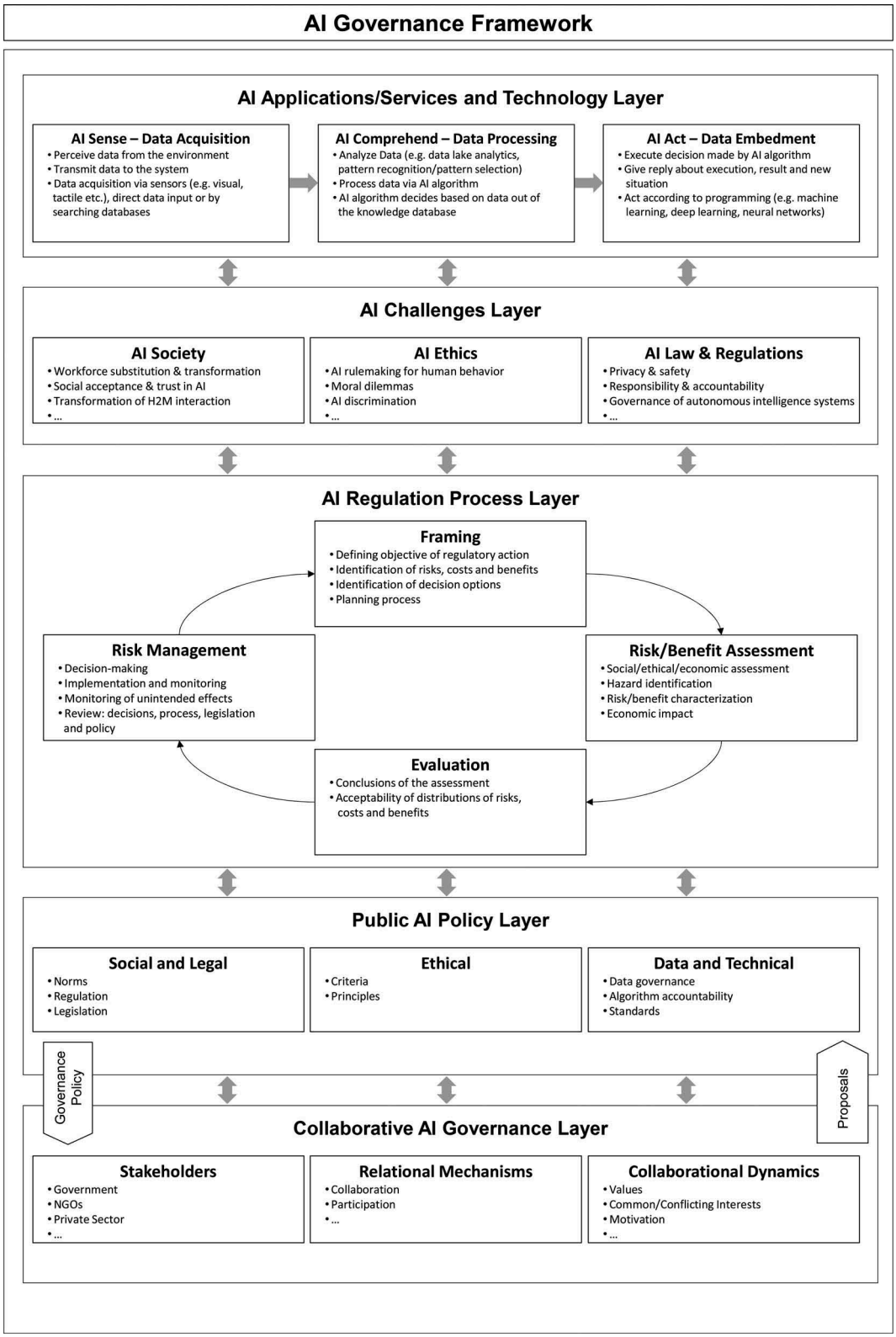

Figure 2. The integrated AI governance framework.

as a whole. To evaluate the success of regulation, it is also necessary to define indicators for measuring the risks and benefits of the planned regulation as well as its effects on different stakeholders. Following this, further actions are required to plan and document the risk assessment as well as the evaluation and risk management to guide future decisions. At the end of this stage, the resources available are distributed and

allocated among the stakeholders involved to enable further actions.

The next step in the regulation process is the *assessment of risks, benefits and costs* itself. Expert groups of the stakeholders collect the data necessary for risk assessment. To provide a full view, a clear definition of the risks and benefits has to be elaborated to identify and measure them in a realistic environment. Details are of great importance, as future regulation has impacts on a social, economic and ethical level. Experimental work or field research are appropriate tools to gather appropriate data and to estimate the possible social and economic impact to obtain a comprehensive and thorough picture of the risks and benefits of AI.

After the assessment is complete, the gathered data are evaluated. In this *evaluation*, the risks and benefits are compared with each other to see who will be affected in which way. In addition, the costs of both potential courses of action, i.e. regulation or no regulation are considered for the stakeholders. Thereafter, the government needs to decide what risks and costs are acceptable and what risks and harmful effects are severe enough to justify interventions via regulation.

The last stage of the process is the *risk management* or in the case of AI, the regulatory action. This involves a critical review of the outcome of the evaluation and the decision for a course of action. This decision is then implemented and the best regulation is enacted and enforced to counter the risks and increase the benefits. The different forms of regulation are discussed in the next layer of the framework. The regulatory process is not finished at this point but rather ends in an evaluation of success of the implementation and potential unintended side effects that could give rise to other challenges. This evaluation is performed by means of indicators developed in the framing stage to evaluate and monitor the results of the regulation. In addition, the regulatory process itself is analyzed to improve future regulation attempts.

This layer represents the concrete outcome of the regulation process. There are different kinds of regulation that may serve as a countermeasure, depending on the area in which the challenges occur. Because of the high uncertainty and the great number of stakeholders involved in the regulatory process, regulation of AI might take a longer time (Scherer, 2016).

According to the governance model of Gasser and Almeida (2017), the instruments to apply regulation can be differentiated based on the time it takes to implement them. For a near-term timeframe, technical challenges can be addressed by introducing or improving industry standards or guidelines to promote the disclosure of AI and data use, as well as explainability and responsibility to prevent discrimination. Similar standards have been already in use for personal care robots since the year 2014 (International Organization for Standardization, 2014).

In the medium-term, more complex, regulatory issues that are ethical in nature can be resolved. Creating an environment of fairness and trust for AI technology and services, requires moral principles and ethical criteria, such as the Asilomar AI Principles (Future of Life Institute, 2017) or the general principles of AI (IEEE, 2017). Such ethical landmarks can guide and evaluate the outcome of the regulatory process and the generation of new and improved AI ethics, providing a mindset for the fair and beneficial development of AI. In the long term, governments can implement new norms and laws to solve social and legal challenges of AI. Especially challenges concerning the workforce changes might require a longer investigation to clearly picture the effects of robots and automation on jobs and employment rates, as well as the effect of possible changes in the legislation.

Collaborative AI governance layer

This layer represents the actors in the regulatory process and strongly relates to the regulatory process and the resulting policies in the policy layer. Many researchers acknowledge that this process not only has to include government that enforces regulation, but also representatives or experts from private organizations, NGOs and agencies (Cath et al., 2017; Scherer, 2016). To balance common and conflicting interests, it is necessary that different interest groups have a shared motivation or shared values of generating beneficial effects with AI technology. A belief, trust and commitment in this idea behind AI is needed to maximize positive effects and acceptance of AI in society (Smith, 2018).

Such collaboration could take many forms, like committees, foundations or agencies. For instance, a new agency consisting of representatives of government and experts from organizations could take over responsibility for the regulatory process as a whole, providing insights into current and future AI issues and making policy proposals to government. This agency could also govern the communication between private and public organizations and propose standards and laws developed with the best knowledge of both the legislative realm and developers or organizations in the field of AI applications (Scherer, 2016; Thierer et al., 2017).

Discussion and conclusion

The recent re-emergence of AI technology with its possibilities and risks has attracted increasing attention, raising the need for governance and regulation. Public administration can hardly keep up with the rapid development of AI, which is reflected in the lack of concrete AI governance and legislation programs. While the challenges of AI and potential adverse effects on society have recently begun to come to attention of researchers, the issue of AI governance and regulation has been widely neglected so far and public administration research has failed to address this matter comprehensively. The analysis of current attempts of governance and regulation of AI in the literature demonstrates that respective strategies of governments and regulatory ideas from private organizations fail to provide a conclusive and concrete way on how to make and implement new policies. Only few regulatory models of AI exist that support the strategic attempts by governments to find a balance between the support of unhindered progression and regulatory control. However, these models lack theoretical foundation and neither consider the challenges and risks of AI as causes of AI governance nor explicitly relate to the context of public administration.

In response to these shortcomings and the need for profound AI governance and regulation concepts, this study proposes an integrated AI governance framework based on regulation theory that depicts the key elements of AI governance and regulation in the context of public administration. The study contributes to public administration research by providing guidance for implementing a comprehensive regulatory process and a frame of reference for future research. The framework helps to structure the heterogeneous and interdisciplinary field of AI and to build the groundwork for promoting its practical use in public administration, as well as its acceptance and beneficial application in society.

Against this background, the study carries several implications for research and practice. The starting point of the framework refers to the challenges and adverse effects associated with AI. While some challenges can be viewed separately, many of them are highly interconnected, imposing social, ethical and legal issues at the same time (Thierer et al., 2017). For example, the use of facial recognition and video surveillance to prevent criminal behavior can reduce crime rates, but also interferes with citizens' privacy and perceived freedom in public spaces (Power, 2016). Careful consideration of such interventions is needed to apply regulatory methods that protect society from respective violations.

Similarly, the main concepts are strongly linked to each other. The framework combines the formerly separately addressed issues of regulation (Beales et al., 2017) and AI challenges (Power, 2016) to provide a holistic view on these strong interconnections. For example, the AI challenges layer is strongly connected to the AI regulation process layer, as the challenges are the very reason regulation is needed. The regulation process itself is linked to the challenges, as each challenge needs a specific approach of regulation to reach the best possible policies and to reap the benefits of AI without allowing it to harm society. Such interdependencies support an integrative approach to AI governance and regulation; as otherwise, they may remain unseen leading to missed benefits or unintended side effects. In contrast to the few other AI governance models (Gasser & Almeida, 2017; Rahwan, 2018), this study also provides a detailed explanation for a regulatory process, in which measures of regulation are developed, evaluated and enacted. This process may serve as a guide and considerate procedure for government and public administration to enact policies in response to the rapid diffusion of AI and its consequences (Stone et al., 2016). While the regulatory process as such strongly focuses on the role of government, the framework also emphasizes the importance of collaborative aspects of different stakeholders that need to work together to deliver the most beneficial outcome to society. Therefore, the actors need a shared motivation or common values to improve the technology and the according rule system. The involvement of different actors secures a balance between public and private interests within this process.

Another important reason for collaborations refers to the imbalanced distribution of knowledge with regard to AI. As AI is a very profitable industry, private organizations have a high personnel requirement and increasingly recruit respective experts to develop new applications and services. The low number of AI experts available leads to high competition for talent among private and public organizations. Due to higher salaries and better job conditions, many AI experts join private organizations and unintentionally produce a knowledge deficit in the public sector, slowing down and aggravating the process of regulation at the same time (Sample, 2017). Therefore, governmental institutions should form collaborations with private organizations to benefit from their advanced knowledge.

Such collaborations already exist and private organizations are supporting regulatory processes. For example, the Information Technology Industry Council (ITI) acts as a representative for many private organizations in the technological field and supports the promotion of

responsible, safe and liable AI technology and explicitly invites governments to form public-private partnerships to improve knowledge transfer and adapt to and prepare for societal changes (ITI, 2017). With the increasing diffusion of AI technology, collaborations might need to overcome national boundaries for regulatory efforts to succeed.

To enact beneficial AI policies, public officers addressing governance and regulation need to be aware of all relevant aspects and must see the "big picture" of AI governance. The framework provides a systematic and comprehensive conceptual guide for public officers dealing with governance-related and regulatory issues of AI. For example, policy makers need to be aware of major technological developments in AI to be able to respond quickly with appropriate policies. Detailed knowledge of every aspect is essential, as new inventions can elicit a great variety of different AI challenges that affect society and public officers themselves (Wirtz et al., 2019). An opportunity to acquire this form of knowledge are exploratory expert interviews with public officials. Furthermore, employing a systematic regulation process can increase the success of the policy outcome, as this approach ensures a comprehensive, thorough and reliable assessment of all different perspectives on the issue in order to determine the best possible regulatory action for public administration. This systematic approach also accelerates the regulatory process by providing a structured plan on which steps are necessary to reach the intended policy.

Forming collaborations is a vital aspect for successfully implementing new policies. To form collaborations within the complex field of AI, it appears reasonable to adapt already established collaborative formats. The field of open innovation, for instance, provides a variety of models and ideas for collaborations, such as ideation platforms (Kaplan & Haenlein, 2010) to gather insights or lead-user approaches (Von Hippel, 2005) to identify the needs of citizens, which can be adapted by public officers to serve the regulatory process of AI.

Despite the above-mentioned contributions, this study is also subject to some limitations, which may represent promising starting points for future research endeavors. While the framework focuses on the challenges and risks of AI as the cause for regulation, it does not explicitly address the potential benefits of AI in the public sector. The assessment of benefits, however, has to focus on the AI, its algorithms, its area of application and affected stakeholders. In the framework, this kind of analysis is part of the evaluation in the regulatory process itself. However, as suggested in the evaluation step of the regulatory process, not only a close inspection of benefits is essential to conclude the usefulness of and the improvements attainable with AI, but also an assessment of the costs and potential tradeoffs associated with regulation (Thierer et al., 2017).

Furthermore, as the integrated conceptual AI governance framework is overarching in nature focusing on the key elements of AI governance and regulation, it is not able to provide concrete guidelines and regulatory measures that can be used to address specific risks or challenges. This matter requires more detailed research within the regulatory process, which could take more time, as some consequences of regulation may be difficult to measure. The further implementation of rules or laws faces similar difficulties, as effects of regulation may be hard to estimate in advance. For example, the increasing replacement of jobs by AI requires the development of a system that supports the now unemployed workers. While the idea of a robot tax for industry or a guaranteed base income for the unemployed might be appropriate (Stone et al., 2016), it will certainly take some time to find regulatory solutions that are fair to all stakeholders.

Another limitation refers to practical problems with the concept of regulation itself. Governmental regulation has always faced the problem of reactivity, providing policies to problems that have already occurred in reality. In addition, the process of regulation can take a very long time due to unclear problems, unprecise definitions and great bureaucratic efforts (Boesl & Bode, 2016). Although the framework provides short-, mid- and long-term policy suggestions, it is impossible to define an exact time frame without knowledge of the specific problem, its effects and all stakeholders involved. Future research could address the question of time by interviewing experts such as lawyers or judges to improve future planning and organization of the regulatory process. Overall, further research is essential to conceptually refine and empirically test the framework developed to improve the inchoate understanding of AI governance and regulation in the public realm.

References

AGCS. (2018). *The rise of artificial intelligence: Future outlook and emerging risks*. Allianz Group. https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/AGCS-Artificial-Intelligence-Outlook-and-Risks.pdf

Ansip, A. (2017). *Making the most of robotics and artificial intelligence in Europe*. European Commission. https://ec.europa.eu/commission/commissioners/2014-2019/ansip/blog/making-most-robotics-and-artificial-intelligence-europe_en

Asaduzzaman, M., & Virtanen, P. (2016). Governance theories and models. In A. Farazmand (Ed.), *Global encyclopedia of public administration, public policy, and governance* (Vol. 65, pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-319-31816-5_2612-1

Asimov, I. (1950). *I, robot*. Gnome Press. https://www.ttu.ee/public/m/mart-murdvee/Techno-Psy/Isaac_Asimov_-_I_Robot.pdf

Baldwin, R., Cave, M., & Lodge, M. (2012). *Understanding regulation: Theory, strategy, and practice* (2nd ed.). Oxford University Press.

Banerjee, S., Singh, P. K., & Bajpai, J. (2017). A comparative study on decision-making capability between human and artificial intelligence. In B. K. Panigrahi, M. N. Hoda, V. Sharma, & S. Goel (Eds.), *Advances in intelligent systems and computing. nature inspired computing* (Vol. 652, pp. 203–210). Springer Berlin Heidelberg. https://doi.org/10.1007/978-981-10-6747-1_23

Bataller, C., & Harris, J. (2016). Turning artificial intelligence into business value. *Today*. https://www.accenture.com/t20160814T215045__w__/us-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Technology_11/Accenture-Turning-Artificial-Intelligence-into-Business-Value.pdf

Beales, H., Brito, J., Davis, J. K., DeMuth, C., Devine, D., Dudley, S., Mannix, B., & McGinnis, J. O. (2017). *Government regulation: The good, the bad, & the ugly, released by the regulatory transparency project of the federalist society*. The Federalist Society. https://regproject.org/wp-content/uploads/RTP-Regulatory-Process-Working-GroupPaper.pdf

Bergen, M. (2018). *Google grapples with 'horrifying' reaction to uncanny AI tech*. Bloomberg. https://www.bloomberg.com/news/articles/2018-05-10/google-grapples-with-horrifying-reaction-to-uncanny-ai-tech

Bleicher, A. (2017). *Demystifying the black box that is ai: Humans are increasingly entrusting our security, health and safety to "black box" intelligent machines*. Scientific American. https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/

Boesl, D. B. O., & Bode, M. (2016). Technology governance. In *2016 IEEE international conference on emerging technologies and innovative business practices for the transformation of societies (EmergiTech)*. IEEE.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge University Press.

Boyd, M., & Wilson, N. (2017). Rapid developments in artificial intelligence: how might the New Zealand government respond? *Policy Quarterly*, *13*(4), 36–44. https://doi.org/10.26686/pq.v13i4.4619

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2017). Artificial intelligence and the 'good society': The US, EU, and UK Approach. *Science and Engineering Ethics*, *24*(2), 505–528. https://doi.org/10.1007/s11948-017-9901-7

Coles, T. (2018). *How GDPR requirements affect AI and data collection*. ITPro Today. http://www.itprotoday.com/risk-and-compliance/how-gdpr-requirements-affect-ai-and-data-collection

COMEST. (2017). *Report of COMEST on robotics ethics*. UNESCO. http://unesdoc.unesco.org/images/0025/002539/253952E.pdf

Cuthbertson, A. (2018). *Elon Musk and Stephen Hawking warn of artificial intelligence arms race*. Newsweek. https://www.newsweek.com/ai-asilomar-principles-artificial-intelligence-elon-musk-550525

De Geest, G. (Ed.). (2017). *Encyclopedia of law and economics* (2nd ed.). Elgar. https://doi.org/10.4337/9781782547457

Delucci, M. A. (2000). Environmental externalities of motor-vehicle use in the US. *Journal of Transport Economics and Policy*, *34* (2), 135–168. http://www.jstor.org/stable/20053837

Den Hertog, J. A. (2012). Economic theories of regulation. In R. van den Bergh & A. M. Pacces (Eds.), *Encyclopedia of law and economics: Vol. 9. Regulation and economics* (Vol. 9, 2nd ed., pp. 25–95). Elgar.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542* (7639), 115–118. https://doi.org/10.1038/nature21056

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Future of Life Institute. (2017). *Asilomar AI principles*. Future of Life Institute. https://futureoflife.org/ai-principles/

Gasser, U., & Almeida, V. (2017). A layered model for AI governance. *IEEE Internet Computing*, *21*(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoeven, J., Zicari, R., Zwitter, A. (2017). Will democracy survive big data and artificial intelligence. *Scientific American*. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

Heyns, C. (2014). *Report of the special rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns*. Human Rights Council of the United Nations General Assembly. https://digitallibrary.un.org/record/771922/files/A_HRC_26_36-EN.pdf

IEEE. (2017). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2*. IEEE. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

International Organization for Standardization. (2014). *ISO 13482:2014: Robots and robotic devices - safety requirements for personal care robots*. International Organization for Standardization. https://www.iso.org/standard/53820.html

ITI. (2017). *Artificial intelligence policy principles*. ITI. https://www.itic.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

König, A., Kuiper, H. A., Marvin, H. J. P., Boon, P. E., Busk, L., Cnudde, F., Cope, S., Davies, H. V., Dreyer, M., Frewer, L. J., Kaiser, M., Kleter, G. A., Knudsen, I., Pascal, G., Prandini, A., Renn, O., Smith, M. R., Traill, B. W., Voet, H. V. D., Vos, E., & Wentholt, M. T. A. (2010). The SAFE FOODS framework for improved risk analysis of foods. *Food Control*, *21*(12), 1566–1587. https://doi.org/10.1016/j.foodcont.2010.02.012

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. San Luis Obispo, CA: California Polytechnic State University.

McGinnis, J. O. (2010). Accelerating AI. *Northwestern University Law Review*, *104*(3), 1253–1269. https://doi.org/10.2139/ssrn.1593851

Mills, T. (2018). *The impact of artificial intelligence in the everyday lives of consumers*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2018/03/07/the-impact-of-artificial-intelligence-in-the-everyday-lives-of-consumers/

Power, D. J. (2016). "Big brother" can watch us. *Journal of Decision Systems*, *25*(51), 578–588. https://doi.org/10.1080/12460125.2016.1187420

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Sample, I. (2017). *Big tech firms' AI hiring frenzy leads to brain drain at UK universities*. The Guardian. https://www.theguardian.com/science/2017/nov/02/big-tech-firms-google-ai-hiring-frenzy-brain-drain-uk-universities

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risk, challenges, competencies, and strategies. *Harvard Journal of Law & Technology*, *29*(2), 354–400. https://dx.doi.org/10.2139/ssrn.2609777

Sirosh, J. (2017). *Delivering AI with data: The next generation of the microsoft data platform*. Microsoft. https://blogs.technet.microsoft.com/dataplatforminsider/2017/04/19/delivering-ai-with-data-the-next-generation-of-microsofts-data-platform/

Smith, B. (2018). *Facial recognition technology: The need for public regulation and corporate responsibility*. Microsoft. https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/

Stemler, A. (2016). Regulation 2.0: The marriage of new governance and lex informatica. *Vanderbilt Journal of Entertainment & Technology Law*, *19*(1), 87–132. https://dx.doi.org/10.2139/ssrn.2746229

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., & Teller, A. (2016). *Artificial intelligence and life in 2030. One hundred year study on artificial intelligence: Report of the 2015-2016 study panel*. Stanford, CA: Stanford University. https://ai100.stanford.edu/2016-report

Thierer, A., O'Sullivan, A., & Russell, R. (2017). *Artificial intelligence and public policy*. Arlington, VA: Mercatus Center George Mason University. https://www.mercatus.org/system/files/thierer-artificial-intelligence-policy-mr-mercatus-v1.pdf

Veale, M., van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Chi (Ed.), *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–14). ACM. https://doi.org/10.1145/3173574.3174014

Von Hippel, E. (2005). *Democratizing Innovation*. MIT Press.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, *2*(6), eaan6080. https://doi.org/10.1126/scirobotics.aan6080

Wang, W., & Siau, K. (2018). Artificial intelligence: A study on governance, policies, and regulations. In *MWAIS 2018 proceedings*. Association for Information Systems. http://aisel.aisnet.org/mwais2018/40

Weyerer, J. C., & Langer, P. F. (2019). Garbage in, garbage out: The vicious cycle of AI-based discrimination in the public sector. In Y.-C. Chen, F. Salem, & A. Zuiderwijk (Eds.), *20th annual international conference on digital government research - dg.o 2019* (pp. 509–511). ACM Press. https://doi.10.1145/3325112.3328220

Wirtz, B., Weyerer, J., & Geyer, C. (2019). Artificial Intelligence and the Public Sector – Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103.

Zheng, Y., Yu, H., Cui, L., Miao, C., Leung, C., & Yang, Q. (2018). SmartHS: An AI platform for improving government service provision. In *32nd AAAI conference on artificial intelligence*. AAAI. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16041/16369