Electronic Markets (2021) 32:59–71
https://doi.org/10.1007/s12525-021-00480-5

RESEARCH PAPER

Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines

Kai Jia[1] · **Nan Zhang**[2]

Received: 31 December 2020 / Accepted: 3 May 2021 / Published online: 2 July 2021

Abstract

Background Governments, enterprises, civil organizations, and academics are engaged to promote normative guidelines aimed at regulating the development and application of Artificial Intelligence (AI) in different fields such as judicial assistance, social governance, and business services.

Aim Although more than 160 guidelines have been proposed globally, it remains uncertain whether they are sufficient to meet the governance challenges of AI. Given the absence of a holistic theoretical framework to analyze the potential risk of AI, it is difficult to determine what is overestimated and what is missing in the extant guidelines. Based on the classic theoretical model in the field of risk management, we developed a four-dimensional structure as a benchmark to analyze the risk of AI and its corresponding governance measures. The structure consists of four pairs of risks: specific-general, legal-ethical, individual-collective and generational-transgenerational.

Method Using the framework, a comparative study of the extant guidelines is conducted by coding the 123 guidelines with 1023 articles.

Result We find that the extant guidelines are eccentric, while collective risk and generational risk are largely underestimated by stakeholders. Based on this analysis, three gaps and conflicts are outlined for future improvements.

Keywords Artificial Intelligence · Risk management · Categorization · Eccentricity

JEL H83

Introduction

Recent years have witnessed the rapid development and application growth of artificial intelligence (AI) across various fields, including transportation, production chains, medical treatment, personal assistance, knowledge production and even political propagation (Harari, 2017; Appenzeller, 2017; Abubakar et al., 2019; Floridi et al., 2020). Based on the powerful transformative force of technological innovation, the dark sides, or negative impacts, of AI have been frequently debated among academicians, policy makers, enterprisers and the public. Fears that AI might marginalize human workers, deteriorate the existing divides, discriminate unprivileged workers and impede privacy have been at the forefront of recent literature and media coverage (Biswas & Mukhopadhyay, 2018; Obermeyer et al., 2019; Turton & Martin, 2020). Rather than drafting hard laws, which are legally binding regulations to define permitted or prohibited conduct, public and private stakeholders have mainly responded to concerns by promoting soft laws in the form of normative ethics guidelines, aiming to constrain the dark sides of AI while preserving innovation (Calo, 2017; Cath et al., 2018). According to the database of AlgorithmWatch, there have been more than 160 AI ethics guidelines globally in the past 5 years.[1]

Responsible Editor: Lin Xiao

✉ Nan Zhang
nanzhang@tsinghua.edu.cn

Kai Jia
jiakai@uestc.edu.cn

[1] School of Public Affairs and Administration, University of Electronic Science and Technology of China, Xi Yuan Street 2006, 611731 Chengdu, China

[2] School of Public Policy and Management, Tsinghua University, 100084 Beijing, China

[1] A comprehensive dataset of global AI Ethic Guidelines *See*: https://inventory.algorithmwatch.org

Despite the merits of soft laws in their agility and flexibility, theoretical and empirical critics have been proposed concerning their effectiveness in guiding the conduct of stakeholders. Researchers doubted the motivations behind these guidelines, especially those supported by private stakeholders, as they may be used as a disguise to either render a social problem technical or discourage the efforts of imposing real regulatory burdens (Benkler, 2019). Using controlled experimental methods, researchers also found that ethical guidelines failed to change the behaviors of professionals from the tech community (McNamara et al., 2018). Beyond the mere negation on the effectiveness of these guidelines, more researchers focus on their contents. They reviewed the existing guidelines and summarized the consensus while analyzing the omissions or conflicts (Greene et al., 2019; Hagendorff, 2020; Jobin et al., 2019).

Given the growing academic interest in rethinking the global landscape of AI ethics guidelines, we still lack a comprehensive understanding of whether they are asking the right question. More clearly, do the concerns included in the guidelines match the AI risks in real life? If they do not, what is the eccentricity of the guidelines, and how can we make improvements in the future? Most of the extant literature subjectively selected a certain number of guidelines and drew conclusions by comparing the contents. However, a holistic framework is still missing to categorize the risk of AI; this framework could be employed to comprehensively and objectively evaluate the existing guidelines. In this paper, we will fill the research gap by scoping studies (Arksey & O'Malley, 2005). Extant literature argues that to design AI for social good, both technical and ethic factors, the latter of which is more connected to the utilization environments, need to be covered (Floridi et al., 2020). Our research on the evaluation work could also be seen as efforts to guide the design of AI from the ethic perspectives.

The following paper is divided into five sections. "Literature review" reviews extant literature on AI risk and the growing interest in the critical study of AI ethics guidelines, illustrating the research gap and lack of a holistic framework on AI risk to evaluate the eccentricity of these guidelines." Theoretical framework of AI risk" proposes the theoretical framework, which consists of four dimensions to categorize the risk of AI. "Methodology" explains the methodology and "Description of the global AI guidelines" provides a descriptive explanation of the global landscape of the guidelines. "Code analysis of the articles" further evaluates the extant 160 + guidelines using the framework proposed in "Theoretical framework of AI risk" and explains the coding results. "Research and managerial implications" and "Conclusions" discusses the policy implications and concludes the paper.

Literature review

Literature on AI risk

Recent years have witnessed the rapid development and application of AI, which is widely heralded as the fourth industrial revolution (Syam & Sharma, 2018). For instance, extant literature utilizes AI to detect the cyber bullying behavior (Sánchez-Medina et al., 2020), predict the performance and turnover acts of employees (Sajjadiani et al., 2019), identify critical hotel cancellations (Sánchez et al., 2020) and its universal application to provide ample public services (Vogl et al., 2020). The transformative impacts of AI can be seen from two perspectives. On the one hand, as Marc Andreessen proposed in 2011, software is dominating the world (Andreessen, 2011). Software, in the form of code, has been the new rule of society that governs the conduct of humans. On the other hand, the importance and effectiveness of the new rule are constrained by the "Polanyi Dilemma", proposed by Michael Polanyi, which indicates that people know more than they can speak (Polanyi, 2009). As software is written by humans, the Polanyi Dilemma means that the process of digitalization is constrained, and there are still many scenes where software could not be utilized. The transformative impacts of AI lie in the break of the Polanyi Dilemma, as it is no longer necessary for humans to develop software; rather, the algorithm could sum up the characteristics based on the big data provided as the input. As a result, the development of AI helps software, or more generally the code, more widely and deeply realize itself as the new rule of human society, echoing the famous slogan of "code is law" proposed by Lawrence Lessig 10 years ago (Lessig, 2009).

From the perspective of rules, the transformative power of AI would be a double-edged sword. Extant literature has proposed that AI creates novel ethical, legal and social challenges (Floridi & Cowls, 2019). Other scholars argued in more detail about specific challenges, three of which are mostly mentioned. First, technical unexplainability, or a "black box", might blur the boundaries of different stakeholders when legal accountability has to be confirmed (Liu et al., 2019). Second, the self-reinforcing mechanism of AI would probably exacerbate existing social problems such as bias, discrimination, echoing chambers, etc. (Nelson, 2019). Third, the subjectivity of AI would cause ethical or legal problems, especially in fields where rights were once only given to humans (Balkin, 2018). Issues such as whether algorithms could be protected by free speech rights and artifacts produced by AI that could be protected by copyright laws are typical and pressing examples.

Despite the ample study of AI risks, most of the literature focuses on specific risks based on case studies.

However, given the characteristic of AI as a general-purpose technology, the risks it provokes are much more comprehensive and systematic than the current literature mentions. Few scholars have developed theoretical frameworks to obtain a holistic view of AI challenges. Recent research has proposed a principle-agent model, which is focused on algorithms supported by AI, to categorize the accountability risk of algorithmic governance (Krafft et al., 2020). Nevertheless, the framework cannot be applied to other risks in addition to accountability. More importantly, the extant literature has substantially disregarded inheriting insights from long-lasting risk management research, which could be a benchmark to develop a theoretical framework to analyze the specific field of AI. This article explores the possible connection between the general risk management literature and the current discussion on AI challenges.

Critical review on AI ethic guidelines

Compared with the lack of a theoretical framework to analyze AI risks, the extant literature has substantially focused on governance responses (Thiebes et al., 2020). First, we have seen multiple calls for beneficial AI (Future of Life Institute, 2017), responsible AI (Chinese National Governance Committee for the New Generation Artificial Intelligence, 2019), trustworthy AI (OECD, 2019), etc., all of which could be considered work trying to define the concept of what kind of AI we truly need (Acemoglu & Restrepo, 2020). Second, ethical AI, or moral machines, became a new trend as academicians and engineers are working together to embed and design ethical rules into algorithms to solve governance challenges at the technical level (Awad et al., 2018). Privacy computing (Hong & Landay, 2004), privacy by design (Schaar, 2010), and transparency computing (Zhang et al., 2017) are promising and typical cases. Third, ethical guidelines and principles are proposed by different stakeholders from different perspectives (Torresen, 2018). Given the early phase of AI applications, global decision makers are concerned that binding requirements might hinder the development of technology, leading them to apply ethical guidelines to motivate stakeholders to be aware of the potential risks and self-regulate.

Due to the prevalence of diverse AI ethics guidelines proposed by different stakeholders, a growing number of critical reviews start to rethink whether these ethics guidelines, or more generally soft laws, are effective. In addition to doubts regarding the incentive of private stakeholders' engagements, as well as a comparative study of the omission or conflicts of the contents, the extant literature also points out the "hidden" shortcomings of the guidelines. First, some scholars criticized the limited number of stakeholders included in the process of drafting these guidelines, especially the inclination to technocrats and males, which would inevitably cause the contents to address moral problems primarily through rational and logic-oriented justice rather than empathic and emotional-oriented ethics (Hagendorff, 2020). Second, some research proposed that the extant literature placed too much emphasis on the first-order ethics that define the contents of specific values while disregarding the second-order ethics that explain the moral background of the values proposed (Greene et al., 2019). As a result, extant guidelines only defined what principles should be followed to govern AI risks without explaining why we need these principles and how they work to improve the potential of AI. Following the trend of focusing on the "hidden" background of extant AI ethics guidelines, we argue that the lack of a holistic framework to analyze AI risks renders the 160 + guidelines similar to "headless flies". Different stakeholders propose diverging principles without understanding what part of the problem they are solving about AI risks and whether the guidelines as a whole could cover all the critical AI risks. The deliberation of the risk framework can be partly connected to the appealing of second-order ethics proposed by Greene et al. (2019), trying to explain how the guidelines would help constrain AI risks. We advance the extant research by proposing a holistic framework and providing a comprehensive evaluation of the existing 160 + guidelines. We summarize the extant literature in Table 1 and illustrate the research gap.

Theoretical framework of AI risk

The analysis of AI risk is the starting point to draft AI ethics guidelines. Only with a clear explanation of what risk AI would generate could the draftsman defend the necessity and effectiveness of the principles. The extant literature usually defines risk from a technical perspective, using the life-cycle model of AI to analyze what risk would arise in each period.[2] For example, in the period of collecting the training data, risk might arise because of the bias buried in the dataset. Similarly, when AI is applied in a business environment, the possibility of misuse or performance evaluation indicators that are not reasonably set might generate potential risks. Although the life-cycle model might be useful for managers to utilize targeted measures to constrain specific risk, it fails to understand risk from an ecosystem perspective, seeing AI embedded into a social structure where both internal factors of AI and external factors of AI would cause different risks (Heckmann et al., 2015). The ecosystem perspective might become more important as the prevalence of AI increases. To compensate for these shortcomings, we developed a new AI risk category framework by tracing back

[2] *See* https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/derisking-ai-by-design-how-to-build-risk-management-into-ai-development

Table 1 Summary of extant literature

Research perspective	Research theme	Major conclusions	Literature examples
Bright Sides of AI	Transformative Power of AI	AI helps software to become the new rule of human society, heralding the fourth industrial revolution	(Lessig, 2009; Andreessen, 2011; Syam & Sharma, 2018)
	Utilization of AI in Specific Environments	AI is good at detection, prediction and identification	(Sánchez-Medina et al., 2020; Sánchez et al., 2020; Vogl et al., 2020)
Dark Sides of AI	General Concern and Specific Challenge of AI	AI would bring about novel ethical, legal and social challenges, including unexplainability, self-reinforcing effects, subjectivity liabilities, etc	(Balkin, 2018; Floridi & Cowls, 2019; Liu et al., 2019; Nelson, 2019)
	AI Risk Category	Accountability risk of AI could be categorized by principle-agent model	(Krafft et al., 2020)
Governance Proposals of AI Risks	New Governance Concepts	Beneficial AI, Responsible AI, Trustworthy AI, etc	(Future of Life Institute, 2017; CNGC, 2019; OECD, 2019)
	Technical Innovation on Moral Machines	Privacy Computing, Privacy by Design, Transparency Computing, etc	(Hong & Landay, 2004; Schaar, 2010; Zhang et al., 2017)
	Ethic Guidelines and Principles	Stakeholders need to be aware of AI risks and self-regulate	(Torresen, 2018)
Critics on Extant Governance Proposals of AI Risks	Doubts on the Effectiveness of Extant Proposals	Incentives of private stakeholders are doubtful. There are omissions or conflicts in the proposals	(McNamara et al., 2018; Benkler, 2019)
	Doubts on the Process to Draft Extant Proposals	Limited stakeholders are included into the drafting process	(Jobin et al., 2019; Hagendorff, 2020)
	Ignorance of the "Hidden" Factors of Extant Proposals	A holistic framework to evaluate the extant proposals is missing	(Greene et al., 2019)

to the classic risk management literature rather than focusing on the technical life-cycle model of AI.

Risk is generally defined as the uncertainty of loss (Rosenbloom, 1972; Williams & Heins, 1985). Following the definition, two important characteristics of risk are summarized by scholars. First, risk is uncertain, as we may never be sure when and how risk will happen. One of the most important reasons for the uncertainty is the development of science and technology. Second, risk is harmful. Risk would make people suffer some loss, huge or small, fatal or inessential. In some environments, the loss could also be understood as an unexpected alteration from preset targets (Ni et al., 2010). Based on the two characteristics, the extant literature generally agreed to breakdown risk into two dimensions, *probability* and *severity* (ISO, 2002; Renfroe & Smith, 2007). From the perspective of risk management, the normative argument for the probability of risk is that it is useful for decision makers to qualitatively distinguish between the most and least urgent risks to choose the optimal statistical decision, which could be better than purely random decision-making (Anthony, 2008). On the other hand, the severity dimension, measuring the consequences of risk, would help managers decide how many resources should be utilized to confront challenges (Ni et al., 2010).

Despite the opinions of some critics, the two-axis structure is widely adopted in risk management research, which could also be helpful when applied as a benchmark to analyze AI risk (Cox et al., 2005). However, given that the risks of AI differ significantly in their categorization, we need to modify the framework to apply in this field (Meek et al., 2016).

Concerning the *probability* of AI risk, we focus on factors that affect the uncertainty of whether and how risk would happen. Two factors concerning the characteristics of AI are important here. Given that the technical capacity of AI is in progress and the application mode of AI is not mature, the uncertainty of AI risk depends on the environment with which we are concerned. This environment could be a specific environment where the demands are clear and the capacity of AI technology is sufficient. Therefore, the risk would be *specific* in the form of the stakeholders involved, causes that could be traced, and the responsibilities that

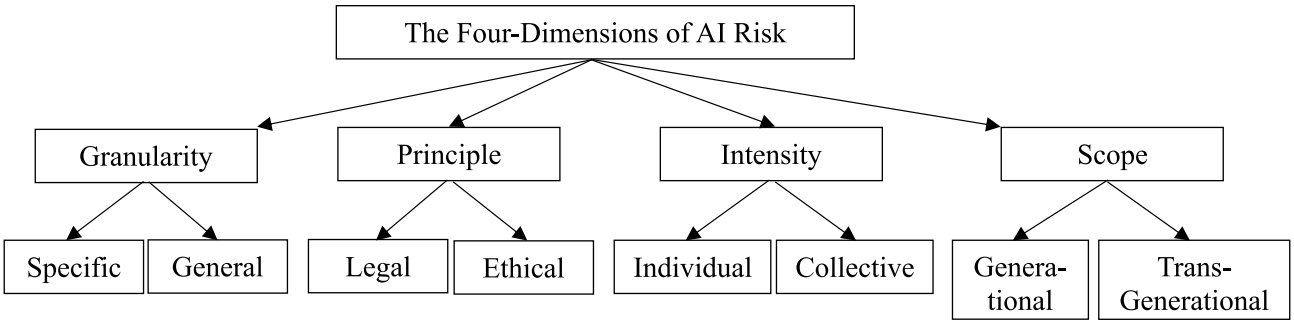

Fig. 1 Theoretical Framework of AI Risk

should be assumed (e.g., discrimination of credit-scoring AI algorithm). The environment could also be a general environment where the potential of AI is recognized but neither the way to apply AI nor the impacts after AI is applied is clear. As a result, the risk would be *general,* which mainly reflects stakeholders' imaginative concern (e.g., rebellion of advanced AI). On the other hand, given the general-purpose technology characteristics of AI, the uncertainty of AI risk also depends on the mechanisms by which AI affects the environment. The effects could be formal, such as by violating the stated rights and interests of humans or organizations, causing *legal* risk. Otherwise, the effects could be informal, e.g., affecting public opinions, which would cause *ethical* risk. Legal risk is more certain than ethical risk, as legal principles are predefined and can be easily recognized when violation happens, while ethical harm might be difficult to discern and usually causes diverging judgments after recognition.

The *severity* of AI risk is usually measured by *Intensity* and *Scope* in the extant literature (Markowski & Mannan, 2008). Intensity describes how many people are affected by risk, either "*individual*" or "*collective*". For example, privacy infringements would be an individual risk, while algorithmic bias would be a collective risk. On the other hand, *Scope* explains how long the risk, either "*generational*" or "*transgenerational*", would last. Concerning risks related to AI, generational risk is more connected with the current technology level and capacity, while transgenerational risk concerns super- or advanced AI, whose risk might extend to future generations.

As a whole, following the classic dichotomy of risk management literature between *probability* and *severity*, we proposed a four-dimensional framework (Fig. 1) to analyze AI risk, each of which includes two categories to distinguish different risks. Although there might be other perspectives to analyze AI risk, we argue that the theoretical framework could cover most of them due to the comprehensiveness of the dichotomy, which has been fully discussed and applied by risk management scholars and practitioners. For example, the extant literature proposed that cultural variations should consider AI ethics guidelines, as Chinese tradition focuses more on group-level equality and social welfare, while European and U.S. approaches prioritize individual autonomy and privacy (Roberts et al., 2020). Despite the differences, the cultural variations could also be explained from the dimension of risk *intensity,* which is under the dimension of risk *severity* in our framework. This finding might support the completeness of our framework. In Table 1, we provide corresponding risk examples for each composition of the four dimensions.

Methodology

In "Theoretical framework of AI risk", we deductively developed a four-dimensional theoretical framework to categorize AI risks that could be employed as a holistic evaluation benchmark to analyze the eccentricity of global AI guidelines. We choose the AlgorithmWatch database as the subjects of the analysis. AlgorithmWatch is a nonprofit organization that was established in Germany. They started the "AI Ethics Guidelines Global Inventory" project in 2019 to compile frameworks and guidelines that seek to set out principles of how AI can be developed and implemented ethically. The inventory has compiled more than 160 guidelines until now, including binding agreements, voluntary commitments, recommendations, etc., but with laws excluded. Given the absence of a unified database for AI ethics guidelines (Jobin et al., 2019), most of the extant literature selected guidelines according to the subjective judgments of authors. To provide a comprehensive evaluation of the global

landscape, we need to compile as many guidelines as possible. The AlgorithmWatch database is suitable for this goal for two reasons. First, the extant literature usually takes the database as the reference to check whether they are missing some guidelines (Hagegndorff, 2020). Second, most of the guidelines mentioned in the extant literature could be found in the database, which at least proves its relative comprehensiveness. For example, most of the guidelines included in the four link hub webpages,[3] which are usually mentioned in the extant literature, can also be found in the AlgorithmWatch database. As the number of guidelines compiled in the database is much larger than others, it is appropriate to use the database as the subjects for the comprehensive evaluation of the global landscape of AI guidelines.

The purpose of the article is to use the theoretical framework to evaluate whether the existing guidelines have covered all the risks, otherwise to illustrate its eccentricity to see what is overstated while others understated. However, different guidelines have diverse structures, most of which propose specific suggestions, while others prefer to analyze without conclusions. To ensure the unified standard of the evaluation, we deleted guidelines that did not have specific articles as conclusions. Additionally, we deleted guidelines that did not have English versions. As a result, 123 guidelines with 1023 articles were selected from the AlgorithmWatch database, forming the subjects of our evaluation. A list of the guidelines and articles are provided with the coding results as the Appendix for further research.

We evaluate each article by manually coding it with four questions according to the risk framework that we proposed in the last section. First, can the risk concerned in the article correspond to a specific reason or general concern? Second, can the risk concerned in the article be controlled using formal legal rules or be guided by values or ideas? Third, are individuals or collectives affected by the risk concerned in the article? Fourth, is the risk concerned in the article the current reality or imagination of the future? To reduce the subjective bias during the coding process, we relied on two group coders who were trained on 10 guidelines with 84 articles to ensure that they had a similar understanding of the evaluation framework. The two group coders independently categorized all 123 guidelines with 1023 articles into four dimensions. For each article in which the two group coders disagreed, the authors discussed the article and formed their conclusions.

[3] *See* Boddington (2018). Alphabetical list of resources. *Ethics for Artificial Intelligence* https://www.cs.ox.ac.uk/efai/resources/alphabetical-list-of-resources/. Winfield (2017). A round up of robotics and AI ethics. *Alan Winfield's Web Log* http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html. National and international AI strategies (2018). *Future of Life Institute* https://futureoflife.org/national-international-ai-strategies. Summaries of AI policy resources. (2018). *Future of Life Institute* https://futureoflife.org/ai-policy-resources/.

Description of the global AI guidelines

A total of 123 AI ethics guidelines were proposed by different stakeholders, of which governments, private sectors, civil society and academia were the top 4, proposing 36, 34, 27, and 14 guidelines with 354, 224, 206, and 95 articles, respectively. Before we perform a detailed analysis of the code, we summarize the most frequently mentioned articles to show the consensus of different stakeholders. In "Code analysis of the articles", we further discuss the eccentricity among the four dimensions. According to our analysis, eight issues are frequently repeated in the guidelines.

Transparency and trust

Transparency is frequently mentioned in most AI guidelines. Given the technical characteristics of the "black box" of AI algorithms, it is often difficult for humans to understand what is safe and what is risky while using AI technology. It might be more difficult to distinguish what is ethical and what is frightening (Goldacre, 2014). If an algorithm is perfect, the "black box" problem would not be that worrisome. However, the empirical consensus in the fields of computer science shows that there are always loopholes and defects in the algorithms (Tan et al., 2014). Transparency is generally considered a necessary prerequisite to reduce the risk of AI. Additionally, it is also considered an important factor to promote public trust and ensure democratic norms during the process of AI development and application.

To meet the requirements of transparency, different standards and norms are proposed in the guidelines. Most of them require the developers and deployers of AI to improve the level of information disclosure, including levels across application mode, source code, data source, etc. (Sampson et al., 2019). Regarding the method of information disclosure, some of the guidelines require interpretation in nontechnical language that can be understood by ordinary people. However, some studies also point out that considering the complexity of AI algorithms, it is difficult for even algorithm designers to provide a clear explanation (Grimmelmann, 2004). Therefore, the concept of trustworthy AI (TAI) is proposed as a new goal. Instead of directly realizing transparency, TAI mainly demands AI to be compliant with all relevant laws and adherent to general ethical principles to make itself perceived as trustworthy by its users (Thiebes et al., 2020). Some research has further argued that beneficence, nonmaleficence, autonomy, justice and explicability should be the top 5 principles to achieve TAI (Floridi et al., 2018), all of which are repeated here.

Bias and equity

Mainstream AI technology is currently based on big data sets and forms rule sets via self-training and self-learning. This technology is a summary of the characteristics of past human social patterns, which can be employed for the perception and decision-making of future activity. AI can improve the efficiency of human society, e.g., the accuracy of advertisement recommendations based on the user's historical preferences. On the other hand, it will be inevitable for AI to mirror the existing divide and differentiation of human society, amplifying biases and exacerbating social equity. In particular, if we consider that the big data set is not necessarily a complete reflection of real society, missing or false data sets may further worsen the fairness problem and eventually lead to systematic discrimination.

Based on these challenges, the vast majority of the guidelines have responded from different angles. Some articles focus on the diversity, inclusiveness and equality of AI training data sets and require AI to ensure the integrity of data sets using a variety of mechanisms in the development process (Microsoft, 2018). Other norms emphasize the results, requiring AI deployers to provide the right to appeal or challenge algorithm decision-making in the application process for affected people, which can "force" AI developers to pay attention to fairness issues. It is worth noting that private sectors are trying to solve the problem of bias and equity via technical means.

Privacy

With the deepening of public awareness of the importance of privacy protection of personal data, an increasing number of guidelines believe that the right of privacy is the basic right of citizens and should be protected in the development and application of AI. Other guidelines connect the protection of privacy with personal freedom and public trust (Bandara et al., 2020), treating it as an important mechanism to prevent the abuse of AI in social monitoring and related fields.

Regarding governance, some guidelines suggest that we should encourage the development of privacy protection technology in the field of AI, of which differential privacy, privacy design and data minimization are typical examples. Some guidelines demand to strengthen the limits on how data are collected and acquired, of which the most important mechanism is to enhance the public's awareness of the data collected by AI. Additionally, some other guidelines demand the protection of personal privacy using government regulations via access control and due processes.

Imputation

The imputation of AI focuses on accountability tracing after the occurrence of risk and the establishment of corresponding remedy systems. Given that the traditional governance system is based on the causality of human behavior, the maturity and popularization of AI inevitably causes a subjectivity problem in the process of imputation. How to divide the civil liability of automatic driving and whether the knowledge products created by AI are protected by copyright laws are typical cases. Some of the guidelines require that the principle of accountability sharing and remedy measures in different situations be specified in advance by contracts. Other guidelines emphasize monitoring the application process of AI to determine the causes of risks and impute them on this basis. To better explore and discover the causes of AI risks, some articles suggest that we should introduce ethical concepts into STEM education and encourage "whistlers" to disclose the potential risks.

Autonomy

Although privacy protection, transparency and other issues are closely related to this issue, how to protect and even promote human freedom and autonomy is still listed as an important concern by many guidelines. Some articles directly propose that the development of AI should aim to enable human freedom and autonomy, indicating that humans should have the right to freely choose whether to use a certain technology (that is, the right to quit the application scenario of AI) and the right to freely choose from different digital platforms or AI technologies (e.g., the right of data portability). Compared with this kind of positive right, some of the articles are relatively conservative and only require that we need to improve the human understanding of AI, especially in the process of data collection and analysis.

Robustness

Many guidelines realize that the risk of AI includes physical harm to people, erosion of social trust (such as "false news" created by AI), disturbance of economic balance, etc. According to these challenges, a variety of solutions are proposed. First, some guidelines require developers and deployers to ensure that there will be no unpredictable risks in AI and establish risk management and control mechanisms. Additionally, some articles focus on solving security risks via technical means. Examples include embedding fairness and security assessments in training data or strengthening the intervention and supervision of the R&D process. Finally, some guidelines also require the innovation of governance mechanisms, including the cooperation of different stakeholders to establish internal supervision and auditing

mechanisms, as well as the third-party evaluation process conducted by industry organizations or user organizations.

Social security

The social security problems caused by AI are concentrated in the field of the labor market. Most of the norms have noticed the impact of AI on the labor market. Although there is a dispute about whether AI will completely replace employment, there is a general consensus that the impact of AI on different labor groups is quite different. Some AI guidelines call for strengthening social security mechanisms, constructing social security networks, actively adjusting the uneven distribution of AI development income, and especially providing social relief for potentially sensitive groups.

In this section, we summarized seven principles that partly illustrate the emphasis of different stakeholders. However, the subjective analysis could not provide any insight to obtain a comprehensive understanding of what is overestimated and what is missing in the extant guidelines. The category of risk and the eccentricity of the propositions still need to be further analyzed from the perspective of a top-down holistic framework, which is proposed in the following two sections.

Code analysis of the articles

In addition to the consensuses that we illustrated in "Description of the global AI guidelines", which were also mentioned in the extant literature (Hagendorff, 2020; Jobin et al., 2019), a more pressing issue is to explain the eccentricity of the existing guidelines to provide a holistic evaluation for future improvements. Based on the theoretical framework proposed in "Theoretical framework of AI risk", we coded the 123 guidelines with 1023 articles from the AlgorithmWatch database.

General analysis of the eccentricity

Of the 1023 articles, 47% of the articles and 60% of the articles were related to specific risks and legal risks, respectively. In addition, 73% and 28% of the articles are related to individual risks and generational risks, respectively. Therefore, it could be inferred that the extant guidelines have a more balanced structure concerning the risk probability and a more eccentric structure related to risk severity.

First, existing guidelines emphasize almost equally the specific risks and general risks when considering the environments where AI challenges occur. The results reflect that AI has been partially utilized in specific fields where causal reasons could be attributed to potential risks. Given that AI is considered to be a general-purpose technology, some general risks still need to be noted beforehand.

Second, legal risks are slightly emphasized compared with ethical risks, indicating that stakeholders are more optimistic that potential challenges of AI could be controlled by formal rules. Nevertheless, 40% of articles focusing on ethical risks show that stakeholders take seriously the new challenges proposed by AI, which cannot be put into traditional legal regimes and must be guided by values or ideas. For the latter, stakeholders need to coordinate with each other to identify proper ways to constrain the risks.

Third, 73% of articles on individual risks indicate that stakeholders believe that most AI impacts would be imposed on individuals rather than collectives. This result is consistent with the current trend of demanding a higher protection level on digital personal rights, especially privacy. However, we can never underestimate the AI impacts on collectives. Actually, the prevalence of AI might be the most complex challenge that we have to face. Consider algorithm bias as an example. It is not the developers who intentionally discriminate specific groups of people with certain identical characteristics but rather the reality that social division is fed back to the algorithms once applied in real environments.

Last, 28% of articles related to generational risks explain that most articles are inclined to concern transgenerational risks that are closely related to super-AI and advanced AI. Although the capacity of AI has been greatly improved, advanced AI is far from possible given the current level of technology development. If we recall the lasting debate on computability in history, whether advanced AI could be realized is still a doubtful question (). Therefore, it might be inappropriate to lead the discussion to unrealistic and even fictional objects, especially when the dark sides of AI have already emerged under some environments, such as facial recognition and smart recommendation. Moreover, the tilt to transgenerational risk could partly be explained because of the undemocratic process of how the guidelines are drafted, as some commenters criticized. Limited by academicians and engineers without the wide participation of the general public, the articles are heavily biased towards the preference of elites.

Analysis of the eccentricity by stakeholders

Different stakeholders have diverging preferences and show different eccentricities on the guidelines. If we divide the eccentricity by the identity of proponents, we can observe their differences in the four dimensions. Table 2 lists the corresponding data of the articles coded. Compared with the total data that we have analyzed, four points deserve to be analyzed in detail. First, governments are more balanced on the scope of AI risks compared with other stakeholders. Forty-one percent of articles proposed by governments focus on generational risk, explaining their relatively realistic attitudes towards the current

Table 2 Four-Dimensions of AI Risk

Granularity dimension	Principle dimension	Intensity dimension	Scope dimension	Example
Specific	Legal	Individual	Generational	Privacy
			Transgenerational	Roberts' Privacy
		Collective	Generational	Self-Driving Car Accident
			Transgenerational	Robot Attack Human
	Ethical	Individual	Generational	Fake Personal Photos
			Transgenerational	Reduced Autonomy
		Collective	Generational	Social Bots Application
			Transgenerational	Job Replacement
General	Legal	Individual	Generational	Copyright Protection of AI Products
			Transgenerational	Robert's Free Speech Right
		Collective	Generational	Discrimination
			Transgenerational	Black Box Effect
	Ethical	Individual	Generational	Human–Machine Relationship
			Transgenerational	Enslaved Humanity
		Collective	Generational	Echoing Chamber Effect
			Transgenerational	Comprehensive Surveillance

development and challenges of AI. Second, private sectors emphasized collective risk relatively more than other stakeholders. This emphasis might partly reflect the advantages that private sectors have about how AI is developed and utilized in real environments, which is challenging for other stakeholders, resulting in their ignorance of collective risk. Third, civil society and academics are inclined to care about individual risk, indicating their emphasis on the protection of personal rights concerning AI challenges. However, this emphasis may also reflect their lack of technical knowledge of how AI risks are formed. Fourth, the very limited concern on generational risk across stakeholders, especially among academia, the private sector and civil society, illustrates a contrasting view that despite the current application of AI, most of the concern on AI risk focuses on the future.

The code analysis clearly shows that extant AI guidelines do not cover all risks evenly but have eccentric distribution. Collective risk and generational risk are substantially underestimated and even ignored by different stakeholders. There might be several reasons to explain the eccentricity. First, as previously mentioned, the process to draft guidelines might limit the range of participators, allowing professionals rather than ordinary users to express their opinions and demands. Second, as the development and application of AI has been in the early stage, some kinds of risks might not be clearly felt and explained, leading to the eccentricity of the framework. Third, the propositions of AI guidelines are affected by public opinion and social cognition about risks, tilting the discussion towards issues that are easy to understand and attracting the most attention. Therefore, it is natural to observe that individual risks, such as privacy, would triumph concern about collective risks, which are more complex and difficult to explain Table 3.

Research and managerial implications

Although the extant literature has gradually realized the dark sides of AI and global stakeholders have proposed numerous AI ethics guidelines trying to constrain the risks, there are an increasing number of critics on the effectiveness of the current work on developing "soft laws" to govern the development and application of AI. Some scholars chose to compare the contents of widely accepted guidelines, which were selected according to their subjective judgments, and draw conclusions on the gaps or conflicts among the guidelines (Jobin et al., 2019). Despite its importance, the reliability of this research is limited because of their methodological shortcomings, as there was no clear standard to select the guidelines, while conclusions might be seriously affected according to the subjective experience of scholars. On the

Table 3 Different stakeholders on four dimensions

	Specific Rrisk	Legal risk	Individual risk	Generational risk
Total	48%	59%	73%	28%
Governments	46%	60%	75%	42%
Private Sector	41%	64%	61%	21%
Civil Society	52%	55%	79%	29%
Academia	41%	57%	88%	9%

other hand, many scholars have started to realize the importance of analyzing the "hidden" factors of extant guidelines. They emphasize more on the stakeholders, process, social and ecological backgrounds of why and how the guidelines are proposed and formed, rather than directly focusing on the contents (Hagendorff, 2020). The research provides a critical theoretical perspective to rethink the necessity and effectiveness of AI ethics guidelines, thus contributing to future improvements in related work. These discussions could provide a foundation for developing a holistic risk analysis framework on AI to promote the evaluation of guidelines and illustrate their focuses and eccentricity.

Prior literature usually proposed an AI risk analysis framework by focusing on the life cycle of AI technology, substantially disregarding the external factors that may also cause risk. We developed a four-dimensional theoretical framework by tracing back to the dichotomy between *probability* and *severity* from the classic risk management literature, thus contributing to the extant research. We admit that this framework is not the only framework to analyze AI risk; nevertheless, we argue that the value of the framework could be seen as a starting point to search for a holistic view to comprehensively evaluate the many guidelines that have emerged globally.

Concerning the coding results, it would be enlightening for future studies to explore why such eccentricity would occur and what impacts it would have on the future development and application of AI. For example, the considerable ignorance of the academic focus on generational risk may partly reflect the limited inclusion of users' experience by scholars, while on the other hand might also ascribe to the social divide between academicians and the public. It might be helpful to explore future studies to empirically test whether and how these factors might systematically affect the outcomes.

Practically speaking, the evaluation of global AI ethics guidelines might provide more implications for future improvements. Based on the eccentric data and related to the contents of the articles, we summarized three points that would be needed more urgently concerning the accelerated speed of technology development and global application of AI.

First, policy makers and managers should be especially aware of the AI risks on collective welfare, rather than assuming that AI would only affect individual rights in the future. According to the coding results, collective risk and generational risk are substantially disregarded by extant guidelines, both of which are related to the understanding of the *severity* of AI risks. The extant literature is inclined to focus more on individual risk and transgenerational risk, resulting in gaps that need to be closed in future work. Academicians and the public might be concerned about individual rights infringements by AI, as individuals are always considered to be less privileged during technology innovation and industrial transformation. The private sector is more motivated by labeling themselves as the protector of consumers rather than on behalf of the collective welfare. However, as general-purpose technology, the transformative power of AI differed from past technologies in its subtleness and comprehensiveness. Critical risks, such as bias and discrimination, are mainly imposed on a group of people rather than individuals. Side effects, such as "Echoing Chamber" or "Social Polarization", are also collective phenomena instead of individual rights infringements. Additionally, all of these risks have emerged because of the prevalent application of AI in fields such as credit scoring and media recommendation, which deserves critical review from regulators and managers.

Second, we need to accelerate the popularization of AI cognitive education so that people can scientifically understand the possible progress and potential risks created by the application of algorithms and form objective expectations while avoiding blind optimism (Awad et al., 2020). As previously mentioned, part of the reason for the eccentricity of the guidelines is that people have not clearly known what the risk would be and how the risk is formed. This question could not be simply explained by researchers or deployers of AI but rather requires coordinated governance behavior from different stakeholders. Even for professional technocrats, obtaining a full understanding of AI risk is a challenging task, which explains why we observe academicians show such an inclination towards transgenerational risk against generational risk. To achieve this purpose, education is a prerequisite to form a cross-disciplinary understanding of AI risk and collectively explore possible governance measures.

Last, global governance regimes on AI urgently need to be developed given the eccentricity of the current isolated and scattering efforts in promoting AI ethics guidelines. According to our analysis, not only are the guidelines as a whole skewed but also the eccentricity differs across stakeholders and countries. Therefore, we need global cooperation to coordinate different stakeholders. Although international organizations such as the UN, OECD, G20, IEEE and WEF could work as dialogue platforms for stakeholders, the more urgently needed efforts are global governance tools and mechanisms. For example, a holistic and widely accepted evaluation framework of the guidelines might be a reasonable starting point, as discussed in this paper. Similarly, universal supervision, punishment and other supporting mechanisms are currently needed.

Conclusions

As artificial intelligence (AI) is enabling significant changes, the dark sides of AI have also been widely recognized globally. Governments, enterprises, civil organizations, and

academics are engaged to promote normative guidelines that aim to regulate the development and application of AI in different fields. Although there have been more than 160 guidelines proposed globally, it is still uncertain whether they are sufficient to meet the governance challenges of AI. Given the absence of a holistic theoretical framework to analyze the potential risk of AI, it is difficult to determine what is overestimated and what is missing in the extant guidelines. This paper proposed a four-dimensional matrix based on the classic theoretical framework of "probability severity" in the field of risk management as a benchmark to analyze the possible risk of AI. The four-dimensional matrix covers four pairs of risks, including specific-general, legal-ethical, individual-collective, and generational-transgenerational risks. Using the framework, a comparative study of the extant guidelines is conducted based on the coding and text mining methodologies. The possible contributions of this paper lie in three points. First, we provide a holistic theoretical framework that could be utilized to analyze the risk of AI. Second, the 123 extant guidelines are coded and mined to illustrate the global concern of different actors. Third, using the four-dimensional framework, a comparative study is conducted to explain the focus and blind spots of extant guidelines, providing suggestions for future research and policies.

Given the contribution, we acknowledge there are also limitations of our research. On one hand, although the AlgorithmWatch dataset is large enough, we may still miss some important guidelines as the process to propose such principles is decentralized and spontaneous. Future research could continue to evaluate the global updates to see whether the eccentricity is getting better or worse. On the other hand, given the purpose of our paper is to provide a holistic view of the global AI ethic guidelines, we have not provided specific suggestions on how the theoretical framework would help specific applications of AI to constrain its possible "dark sides", for which future study could be conducted.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s12525-021-00480-5.

Acknowledgements The authors would like to present thank Meiyin Huang, Peifen Li, Zongyang Li, Sisi Tang, Yu Wang, Haiming Wu, Qianwei Xu, Jing Zhang, who are students and research assistants in the Tsinghua University, and Yanglan Xu, Hao Wang, Zhihao Chen, Mingyang Wei, Jiayu Zhang,who are students in the University of Electronic Science and Technology of China, for their excellent work in the cross-checking of article codes. This work was partially supported by the National Key Research and Development Program of China (2018YFC0832305), the National Natural Science Foundation of China (71974111, 91646103), and the National Social Science Fund of China (18CZZ025).

References

Abubakar, A. M., Behravesh, E., Rezapouraghdam, H., & Yildiz, S. B. (2019). Applying artificial intelligence technique to predict knowledge hiding behavior. *International Journal of Information Management, 49*, 45–57. https://doi.org/10.1016/j.ijinfomgt.2019.02.006

Acemoglu, D., & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society, 13*(1), 25–35. https://doi.org/10.1093/cjres/rsz022

Andreessen, M. (2011). Why software is eating the world. *Wall Street Journal, 20*(2011), C2.

Anthony (Tony) Cox Jr, L. (2008). What's wrong with risk matrices? *Risk Analysis: an International Journal, 28*(2), 497–512. https://doi.org/10.1111/j.1539-6924.2008.01030.x

Appenzeller, T. (2017). *The AI revolution in science*. Science. https://www.sciencemag.org/news/2017/07/ai-revolution-science

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, pp. 19–32. https://doi.org/10.1080/1364557032000119616

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., & Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Awad, E., Anderson, M., Anderson, S. L., & Liao, B. (2020). An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence, 287*, 103349. https://doi.org/10.1016/j.artint.2020.103349

Balkin, J. M. (2018). Free Speech is a Triangle. *Columbia Law Review, 118*(7), 2011–2056.

Bandara, R., Fernando, M., & Akter, S. (2020). Privacy concerns in E-commerce: A taxonomy and a future research agenda. *Electronic Markets*, *30*(3) 629–647. https://doi.org/10.1007/s12525-019-00375-6

Benkler, Y. (2019). Don't let industry write the rules for AI. *Nature, 569*, 161.

Biswas, B., & Mukhopadhyay, A. (2018). G-RAM framework for software risk assessment and mitigation strategies in organizations. *Journal of Enterprise Information Management, 31*(2), 276–299. https://doi.org/10.1108/JEIM-05-2017-0069

Boddington, P. (2018). *Alphabetical list of resources*. Ethics for Artificial Intelligence. https://www.cs.ox.ac.uk/efai/resources/alphabetical-list-of-resources/

Calo, R. (2017). Artificial Intelligence policy: a primer and roadmap. *UCDL Review, 51,* 399.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). 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

Chinese National Governance Committee for the New Generation Artificial Intelligence. (2019). *Governance Principles for the New Generation Artificial Intelligence–Developing Responsible Artificial Intelligence.* China Daily. https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html

Cox, L. A., Jr., Babayev, D., & Huber, W. (2005). Some limitations of qualitative risk rating systems. *Risk Analysis: an International Journal, 25*(3), 651–662. https://doi.org/10.1111/j.1539-6924.2005.00615.x

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: seven essential factors. *Science and Engineering Ethics, 26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

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

Goldacre, B. (2014). *When data gets creepy: the secrets we don't realize we're giving away*. The Guardian. https://www.theguardian.com/technology/2014/dec/05/when-data-gets-creepy-secrets-were-giving-away

Greene, D., Hoffman, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. *Hawaii International Conference on System Sciences (HICSS),* 1–10. https://doi.org/10.24251/HICSS.2019.258

Grimmelmann, J. (2004). Regulation by Software. *Yale LJ, 114*, 1719.

Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. *Minds and Machines*, 1–22. https://doi.org/10.1007/s11023-020-09517-8

Harari, Y. N. (2017). Reboot for the AI revolution. *Nature, 550*, 324–327. https://doi.org/10.1038/550324a

Heckmann, I., Comes, T., & Nickel, S. (2015). A critical review on supply chain risk—definition, measure and modeling, *Omega, 52,* 119–132. https://doi.org/10.1016/j.omega.2014.10.004

Hong, J. I., & Landay, J. A. (2004). An architecture for privacy-sensitive ubiquitous computing. *Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services*, 177–189. https://doi.org/10.1145/990064.990087

ISO. (2002). *Risk Management: Guidelines for use in standards*. ISO/IEC Guide 73.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Krafft, T. D., Zweig, K. A., & König, P. D. (2020). How to regulate algorithmic decision-making: a framework of regulatory requirements for different applications. *Regulation & Governance*. https://doi.org/10.1111/rego.12369

Lessig, L. (2009). Code: And other laws of cyberspace.Version 2.0. New York: Basic Books.

Liu, H. W., Lin, C. F., & Chen, Y. J. (2019). Beyond State v Loomis: artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology, 27*(2), 122–141. https://doi.org/10.1093/ijlit/eaz001

Markowski, A. S., & Mannan, M. S. (2008). Fuzzy risk matrix. *Journal of Hazardous Materials, 159*(1), 152–157. https://doi.org/10.1016/j.jhazmat.2008.03.055

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? In G. T. Leavens, A. Garcia, C. S. Păsăreanu (Eds.) *Proceedings of the 26th ACM joint meeting on european software engineering conference and sym- posium on the foundations of software engineering—ESEC/FSE 2018,* 1–7. New York: ACM Press. https://doi.org/10.1145/3236024.3264833

Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2016). Managing the ethical and risk implications of rapid advances in Artificial Intelligence. *International Conference on Management of Engineering and Technology (PICMET), Portland*, 682–693, 108. https://doi.org/10.1109/PICMET.2016.7806752

Microsoft. (2018). *Responsible bots: 10 guidelines for developers of conversational AI*. https://www.microsoft.com/en-us/research/publication/responsible-bots/

National and international AI strategies. (2018). Future of Life Institute. https://futureoflife.org/national-international-ai-strategies

Nelson, G. S. (2019). Bias in Artificial Intelligence. *North Carolina Medical Journal, 80*(4), 220–222. https://doi.org/10.18043/ncm.80.4.220

Ni, H., Chen, A., & Chen, N. (2010). Some extensions on risk matric approach. *Safety Science, 48*, 1269–1278. https://doi.org/10.1016/j.ssci.2010.04.005

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

OECD. (2019). *OECD Principles on AI.* https://www.oecd.org/going-digital/ai/principles/

Polanyi, M. (2009). *The tacit dimension*. University of Chicago Press.

Renfroe, N. A., & Smith, J. L. (2007). *Whole building design guide: threat/vulnerability assessments and risk analysis*. Washington, DC: National Institute of Building Sciences. http://www.wbdg.org/design/riskanalysis.php

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2020). The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *AI & Society*, 1–19. https://doi.org/10.1007/s00146-020-00992-2

Rosenbloom, J.S. (1972). Case Study in Risk Management. *Prentice Hall,* 63–67.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology, 104*(10), 1207. https://doi.org/10.1037/apl0000405

Sampson, C. J., Arnold, R., Bryan, S., Clarke, P., Ekins, S., Hatswell, A., Hawkins, N., Langham, S., Marshall, D., Sadatsafavi, M., Sullivan, W., Wilson, E. C. F., & Wrightson, T. (2019). Transparency in decision modelling: what, why, who and how?. *Pharmacoeconomics*, 1–15. https://doi.org/10.1007/s40273-019-00819-z

Sánchez, E. C., Sánchez-Medina, A. J., & Pellejero, M. (2020). Identifying critical hotel cancellations using artificial intelligence. *Tourism Management Perspectives, 35*, 100718. https://doi.org/10.1016/j.tmp.2020.100718

Sánchez-Medina, A. J., Galván-Sánchez, I., & Fernández-Monroy, M. (2020). Applying artificial intelligence to explore sexual cyberbullying behaviour. *Heliyon, 6*(1), e03218. https://doi.org/10.1016/j.heliyon.2020.e03218

Schaar, P. (2010). Privacy by design. *Identity in the Information Society, 3*(2), 267–274. https://doi.org/10.1007/s12394-010-0055-x

Summaries of AI policy resources. (2018). Future of Life Institute. https://futureoflife.org/ai-policy-resources/

Syam, N., & Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: machine learning and Artificial Intelligence in sales research and practice. *Industrial Marketing Management, 69*, 135–146. https://doi.org/10.1016/j.indmarman.2017.12.019.

Tan, L., Liu, C., Li, Z., Wang, X., Zhou, Y., & Zhai, C. (2014). Bug characteristics in open source software. *Empirical Software Engineering, 19*(6), 1665–1705. https://doi.org/10.1007/s10664-013-9258-8

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets*, 1–18. https://doi.org/10.1007/s12525-020-00441-4

Torresen, J. (2018). A review of future and ethical perspectives of robotics and AI. *Frontiers in Robotics and AI, 4*, 75. https://doi.org/10.3389/frobt.2017.00075

Turton, W., & Martin, A. (2020). *How deepfakes make disinformation more real than ever*. Bloomberg. https://www.bloomberg.com/news/articles/2020-01-06/how-deepfakes-make-disinformation-more-real-than-ever-quicktake

Vogl, T. M., Seidelin, C., Ganesh, B., & Bright, J. (2020). Smart technology and the emergence of algorithmic bureaucracy: Artificial Intelligence in UK local authorities. *Public Administration Review, 80*(6), 946–961. https://doi.org/10.1111/puar.13286

Williams, C. A., & Heins, R. M. (1985). *Risk Management and Insurance,* 7–9. McGraw Hill.

Winfield, A. (2017). *A round up of robotics and AI ethics*. Alan Winfield's Web Log. http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical

Zhang, Y., Guo, K., Ren, J., Zhou, Y., Wang, J., & Chen, J. (2017). Transparent computing: A promising network computing paradigm. *Computing in Science & Engineering, 19*(1), 7–20. https://doi.org/10.1109/MCSE.2017.17