Applied Artificial Intelligence

An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines

Gaelle Cachat-Rosset & Alain Klarsfeld

APPLIED ARTIFICIAL INTELLIGENCE
2023, VOL. 37, NO. 1, e2176618 (744 pages)
https://doi.org/10.1080/08839514.2023.2176618

Taylor & Francis
Taylor & Francis Group

Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines

Gaelle Cachat-Rosset[a] and Alain Klarsfeld[b]

[a]Department of Management, Faculty of Business Administration, Université Laval, Québec, QC, Canada; [b]Head of Laboratory Work, employment and Health, Toulouse Business School Toulouse, Toulouse, France

ABSTRACT
Artificial intelligence (AI) is present everywhere in the lives of individuals. Unfortunately, several cases of discrimination by AI systems have already been reported. Scholars have warned on risks of AI reproducing existing inequalities or even amplifying them. To tackle these risks and promote responsible AI, many ethics guidelines for AI have emerged recently, including diversity, equity, and inclusion (DEI) principles and practices. However, little is known about the DEI content of these guidelines, and to what extent they meet the most relevant accumulated knowledge from DEI literature. We performed a semi-systematic literature review of the AI guidelines regarding DEI stakes and analyzed 46 guidelines published from 2015 to today. We fleshed out the 14 DEI principles and the 18 DEI practices recommended underlying these 46 guidelines. We found that the guidelines mostly encourage one of the DEI management paradigms, namely fairness, justice, and nondiscrimination, in a limited compliance approach. We found that narrow technical practices are favored over holistic ones. Finally, we conclude that recommended practices for implementing DEI principles in AI should include actions aimed at directly influencing AI actors' behaviors and awareness of DEI risks, rather than just stating intentions and programs.

ARTICLE HISTORY
Received 27 September 2022
Revised 27 January 2023
Accepted 31 January 2023

Introduction

Artificial intelligence (AI) is now present everywhere in the lives of individuals. But only recently has the issue of discrimination by AI been brought to light (O'neil 2016; Zou and Schiebinger 2018). Given this ubiquity, it is essential to ensure that AI systems (AISs) lead to fair and nondiscriminatory decisions. An AIS is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments, with varying levels of autonomy (OECD 2019). At the very least, they risk reproducing existing inequalities or even amplifying them (Bolukbasi et al. 2016; Zou and Schiebinger 2018).

CONTACT Gaelle Cachat-Rosset gaelle.cachat-rosset@fsa.ulaval.ca Department of Management, Faculty of Business Administration, Université Laval, Pavillon Palasis-Prince, 2325 rue de la Terrasse, Québec, QC, Canada, G1V 0A6

The risk of discrimination and unfair treatment in AI mainly stems from two major causes. On the one hand, biased AI learning databases (Eubanks 2018), given the use of shared and reusable datasets, tend to reproduce and maintain initially discriminatory algorithms (Zhao et al. 2017). On the other hand, unconscious biases and stereotypes of AI designers, developers, and trainers who project their own representation of reality or society into their work, can cause discriminatory behaviors from the AISs they develop (West, Whittaker, and Crawford 2019).

Several initiatives have recently been undertaken for fair AISs, such as the development of more diverse AI learning databases (Holstein et al. 2019), and the enactment of principles and guides encompassing ethics and principles for diversity, equity, and inclusion (DEI) for the development, deployment, acquisition, use, and governance of fair and trustworthy AI (European Commission 2020). The recent rise of these comprehensive ethical guidelines for AI, and recommendations for application, showcases the sense of urgency and the strong awareness the international community has toward regulating AI practices for DEI, and avoiding negative impacts on beneficiaries, particularly communities at higher risk of inequity such as women and racial minorities. Nevertheless, little is known on what these principles and guidelines concretely comprise and operationally recommend for DEI specifically.

The Ethical Imperative in AI

As occasions for ethics to be compromised are particularly high for artificial intelligence (owing to its autonomy compared to other artifacts such as medical treatment, the use of which is not traditionally implemented without human oversight), ethical guidelines for artificial intelligence have been proliferating in the last 10 years. Previous reviews (Hagendorff 2020; Jobin, Ienca, and Vayena 2019) were conducted on AI ethics guidelines, and these include a large scope of notions. We will review the main ethics concerns revealed by these previous reviews: morality non-maleficence, autonomy, privacy, accountability, transparency, and fairness.

AI ethics indeed involve a wide range of issues. Among the first, comes the morality of AI-based decisions. Can AISs harm? How does AISs face moral dilemmas? If an AI-powered self-driving car is to choose between killing pedestrians, or putting itself and its driver at risk, what will or should the AIS do? (Nyholm and Smids 2016). Another concern is the autonomy of AISs and the harnessing of some forms of AISs' supposed propensity to break free from human control. By allowing an AIS to learn without supervision, some can venture into unexpected and potentially dangerous directions. In a recent experiment, Facebook had to discontinue two chatbots equipped with machine-learning language capabilities from interacting as they started to develop a language of their own that no one was able to understand (Bradley

2017). Privacy is another domain where AI may be programmed to infringe basic human rights (Zuboff 2019). For instance, legislation in countries such as the USA offers little protection against the use of personal data, particularly with respect to the automated processing of data relating to online activity. Some algorithms originally programmed to improve user experience without compromising user privacy were later amended to allow mining users' search behavior, not for service improvement, but for the sake of market segmentation and targeted advertising. Moreover, as AI takes decision-making powers, such as when to press the brakes (for a driverless car) or whether to allow a loan to a potential borrower, or how to sentence a criminal offender, this raises the question of accountability. If a decision causes damage to some user, defendant or stakeholder, who is to be held responsible for the decision? Should humans relinquish all forms of accountability? (Nyholm and Smids 2016). Linked to accountability, is the transparency of AI-based decision-making systems. AI, particularly when it involves deep learning, may become opaque in the absence of deliberate efforts to monitor its functioning (Larsson and Heintz 2020). Last but not least, the AI ethics literature pays due attention to issues of justice and fairness. For instance, do AI-aided sentences inflict harsher sentences on members of minorities? Do facial recognition algorithms used in recruitment screen out a disproportionate amount of black or colored applicants? Do AI-powered voice recognition systems consider all English accents on an equal basis?

Despite the strong potential impacts of AI on DEI outcomes, to the best of our knowledge, no literature review was dedicated to the topic of exploring DEI in the AI field. This study seeks to fill this gap, by investigating the DEI content of AI ethical guidelines' *principles and practices*, while adopting a DEI management lens. In doing so, we draw on DEI established frameworks, generally ignored in the AI ethical guidelines production and literature. Our first objective is therefore to assess to what extent AI ethical guidelines' *principles* fully cover the mainstream theoretical paradigms for managing diversity widely supported in the DEI literature (Dwertmann, Nishii, and van Knippenberg 2016; Thomas and Ely 1996). Our second objective is to assess whether ethical guidelines' recommended practices actually target AI actors' behaviors.

DEI Stakes in AI Ethical Principles

In his review of 22 guidelines for ethics in AI, Hagendorff (2020) revealed that 80% identified the notion of fairness, situating it as part of the minimum required for the development and use of "ethically sound" AISs. While present in AI ethics guidelines. Fairness is also a central concept in DEI literature, most often understood as the equity dimension of DEI. Jobin, Ienca, and Vayena (2019) analyzed 84 documents containing ethical principles for AI and highlighted 11 principles that mainly converge in ethical guidelines for AI, including one labeled “justice, fairness and equity” found in 81% of the studied documents. But they noticed that the understanding of this principle was not homogenous: if mitigating bias in AI is a common interpretation of fairness, the prevention of discrimination is significantly less referenced by private ethical guidelines, whereas guidelines from the public sector emphasize the impacts of AI on the labor market and the need to address social issues.

DEI Stakes in AI Ethical Recommended Practices

Principles of conduct in the technology sector do not automatically translate into practice (Van den Bergh and Deschoolmeester 2010) and these principles and guidelines are still not widely available to designers (Garcia-Gathright and Springer 2018). McNamara, Smith, and Murphy-Hill (2018) highlighted that the existence of ethical principles alone was inconclusive in influencing designers’ practices in AI. They showed that providing ethical guidelines for decision-making in ethical scenarios to professionals in the technology industry does not change behaviors, as compared to those who did not receive an ethics code. More nuanced findings from Van den Bergh and Deschoolmeester (2010) state that providing an ethics code to ICT professionals may lead them to make more ethical decisions, but this was not true for the “Fairness and Discrimination” situation provided. Mittelstadt (2019) also states that ethics in AI do not benefit from proven methods to translate principles into practice nor from common professional norms and robust legal accountability mechanisms for the AI sector, contrary to the medical sector. Actually, most tools and methods for implementing ethical principles are not so easy to use and do not provide sufficient practical support (Morley et al. 2019). We seek in this study to identify the DEI practices recommended by AI guidelines in the AIS development process and in AIS design organizations, and to what extent AI DEI guidelines may effectively target AI designers’ decisions and behaviors.

Theoretical Insights of DEI Management Research for AI Ethical Guidelines

Principles of justice, fairness, and equity widely promoted in AI guidelines (Hagendorff 2020; Jobin, Ienca, and Vayena 2019), are part of a broader stream of research on DEI. Justice, fairness, and equity refer to a paradigm for adopting and implementing DEI policies and programs in organizations. DEI paradigms are values, beliefs, and norms regarding the reasons and the means to go about diversity management (Kulik 2014). Thomas and Ely (1996) outlined three paradigms when managing DEI: (1) *Discrimination and Fairness*, referring to legal compliance based on efforts to recruit under-represented groups, ensure fairness in organizational treatments and avoid discrimination; (2) *Access and Legitimacy* referring to leveraging employee diversity to better understand and serve a diverse customer base; (3) *Learning and Effectiveness* referring to the use of workforce diversity as a lever for organizational learning through the interaction of actors who have different points of view. In sum, mainstream diversity management theoretical frameworks can be grouped along a distinction between a "fair representation and treatment perspective" on the one hand (1), and a "valuing diversity perspective" on the other hand (2 and 3), that Dwertmann, Nishii, and van Knippenberg (2016) call *Synergy perspective*. While the first perspective is useful in reducing the injustice suffered by traditionally disenfranchised groups as well as legal liability, it is criticized for the lack of acknowledgment of the positive contribution of the members of said groups can make to their employing organizations (Thomas and Ely 1996). The valuing diversity perspective precisely stresses these potential contributions. However, it may be criticized for weakening the moral case put forward by the fair representation and treatment perspective (Lorbiecki and Jack 2000). While often opposed by their originators, these approaches have been convincingly argued as being mainly compatible and complementary (Oswick and Noon 2014). Both might therefore be seen as relevant to AI ethics. Indeed, the primary motivation for embracing a DEI perspective is key to understand the designed and implemented policies and programs that will support the targeted perspective and send signals of expected and rewarded behaviors. So, DEI research offers a broad view of reasons and means to achieve fair and inclusive environments. As such, it is a relevant framework to assess the DEI content of the principles of proposed in guidelines for AI, considering the plethora of DEI approaches in organizations.

Moreover, the DEI literature highlighted that diversity, inclusion, and equity are social constructs for perceiving and judging differences between individuals (Syed and Özbilgin 2009). Diversity is therefore not a universal concept, contrary to the way it was first considered when emerging in the USA in relation to the place of women and racial minorities in the society. This is why DEI management has been adapted to different national and social contexts, in its definitions, targeted groups and management methods, when it spread in Europe in the 2000s and then elsewhere in the world. In particular, the notion of race, which is widely used in the USA, has no agreed biological definition and therefore is not recognized and used in European contexts. For example, Boxenbaum (2006) explains how diversity management has been adapted to the Danish context, which is characterized by strong egalitarianism, meeting both financial and human development expectations. More broadly, comparative studies across countries present the different approaches and perspectives of diversity management at work, showing the cultural and historical nature of the concept (Klarsfeld et al. 2014, 2019). DEI is thus context-sensitive, and it is inappropriate to transpose DEI management principles and practices from one cultural context to another without prior consideration, what Selbst et al. (2019) call the *portability gap* of fairness in AI. In the specific context of AI, Kiemde and Kora (2022, 1) warned that ethics guidelines are dominated by Western works and the contribution of Africa in the literature of AI ethics is very weak, and that "*the predominance of Western input on AI ethics guidelines can lead to a dominance of Western values and vision on AI ethics*." While most of the AISs deployed in Africa comes from Western or Chinese technology giants, they call for the definition of African AI values and aligning AI frameworks with these values (Kiemde and Kora 2022).

Based on the above review, we can formulate our first research questions pertaining to AI ethics guidelines' principles:

> *Research question 1a: What are the DEI principles put forward in AI ethics guidelines?*

> *Research question 1b: Do DEI principles in AI meet the DEI management theoretical paradigms and the DEI context-sensitive approach?*

In addition, Kulik (2014) presents DEI management system components, offering the required structure for enhancing DEI within organizations. In this five-component system of diversity management, the diversity paradigm is declined in policies, operationalized through diversity programs, which in turn orient diversity practices. Finally, the DEI management system is reflected in the incumbents' shared perception of the organizational diversity climate. Thus, diversity climate synthesizes DEI management's impact on employee perceptions. A positive DEI climate has been shown to strengthen positive attitudes and job performance of organizational incumbents (Cachat-rosset, Carillo, and Klarsfeld 2021). 2019) offered a comprehensive conceptualization of diversity climate consisting in three dimensions: intentionality, programming, and praxis. Intentionality refers to employees' perception that the organization is committed to and values diversity; the programming dimension refers to the perception of formal DEI programs and policies set up within the organization to develop and support equity, workforce diversity and global inclusion; and the praxis dimension refers to perceptions about managers' and colleagues' attitudes and behaviors toward diverse people. Each of the three dimensions reflects perceptions of signals sent by the organization, through the voice of its top management, formal programs, and behaviors performed by managers and colleagues. When widely perceived and interpreted by organizational incumbents, these messages create a strong climate (Bowen and Ostroff 2004), that would lead them to adopt the desired behaviors toward DEI. DEI climate has been shown to predict the adoption of pro-DEI behaviors by members of an organization (Chung et al. 2015; Singh, Winkel, and Selvarajan 2013). DEI climate is a relevant concept for AI ethics, as AISs and AIS design organizations generate perceptions among both their designers and users. AI may actually have the potential to expand the diversity climate notion beyond members of organizations, to societies at large, given the pervasiveness of AI. Thus, the three-dimensional conceptualization of diversity climate (Cachat-rosset, Carillo, and Klarsfeld 2019) is yet a relevant lens to assess whether DEI recommended practices in AI guidelines are equipped to influence AI actors' behaviors regarding DEI. We can now formulate our second research questions pertaining AI ethics guideline's practices:

> Research question 2a: What are the recommended practices for the application of DEI principles in AI ethics guidelines?
>
> Research question 2b: To what extent do DEI recommended practices in AI ethics guidelines effectively target DEI behaviors of AI actors?

Methods and Sample

In the current research, our aim was to perform a semi-systematic literature review of DEI principles and practices in AI guidelines. A semi-systematic review is relevant when covering a broad topic emerging from different diverse disciplines and different types of documents, and for detecting themes, perspectives, or common issues within a specific topic (Snyder 2019). It allows to map and analyze the current corpus of principles and practices on DEI in AI. Following semi-systematic literature review guidelines, a rigorous source selection process was implemented (Tranfield, Denyer, and Smart 2003).

Sources Selection

We developed a protocol for selecting eligible sources, adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Moher et al. 2009).

Relevant documents were retrieved following a two-step search strategy (Figure 1). First, we relied on the most comprehensive literature reviews available on ethics guidelines in AI including those regarding DEI (see Hagendorff 2020; Jobin, Ienca, and Vayena 2019). Second, we performed a keyword-based web search. The keywords used for our web search were "artificial intelligence" (and variant "AI"), and "principle/code/guideline/ethic/responsible/trust," and "equality/equity/diversity/inclusion/fairness/justice/discrimination/gender/minority." In both steps, the inclusion criteria were (1) documents published in 2015 and later to focus only on the most recent sources; (2) reliable sources with a large scope of influence, namely governmental or intergovernmental reports (such as the COMEST, the World Economic Forum or the G7), reports from professional communities or

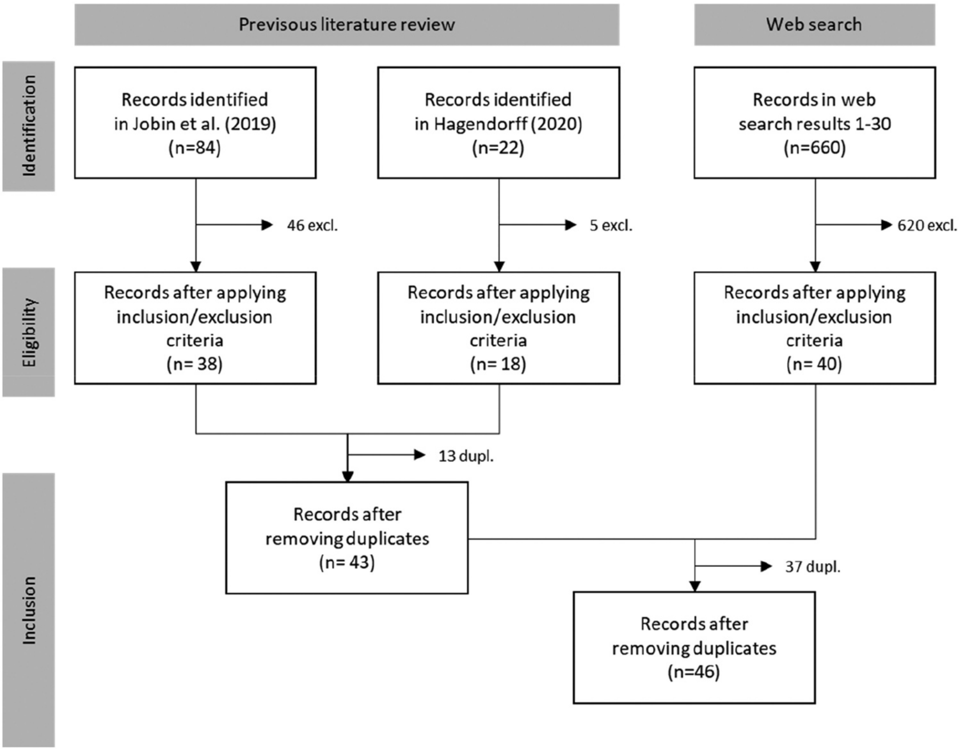
Previsous literature review
Web search
Identification
Eligibility
Inclusion
Records identified in Jobin et al. (2019) (n=84)
Records identified in Hagendorff (2020) (n=22)
Records in web search results 1-30 (n=660)
46 excl.
5 excl.
620 excl.
Records after applying inclusion/exclusion criteria (n= 38)
Records after applying inclusion/exclusion criteria (n= 18)
Records after applying inclusion/exclusion criteria (n= 40)
13 dupl.
Records after removing duplicates (n= 43)
37 dupl.
Records after removing duplicates (n=46)

Figure 1. PRISMA-based flowchart of document retrieval process.

association related to AI (e.g. the Japanese Society for Artificial Intelligence, AI for Humanity, AI4People, AI Now Institute, the Institute of Electrical and Electronics Engineers, Incorporated "IEEE") and reports from influential private companies in AI development (e.g. Microsoft Corporation, Google, IBM or Accenture); (3) documents written in English. We excluded documents not on the subject of AI and company-specific guidelines without larger scope. Finally, duplicates were discarded from the selection. The final sample consisted of 46 documents (Table 1). Eighty percent were published between 2017 and 2019, 26% of sources were of international scope, 35% from North America, 28% from Europe, 9% from Asia and 2% from Australia.

Content Extraction and Categorization

We followed a multi-step categorization strategy, using a deductive categorization based on our expertise on DEI stakes and management, and then an inductive identification of non-preliminarily detected categories. Regarding DEI principles, the deductive categorization yielded 11 categories, while the inductive one brought up 3 additional categories: "human dignity," "remedies for discrimination" and "civil society interaction/inclusion." None of the

Table 1. List of sources (*N* = 46).

Year	Title	Source	Type of document	Geographical scope/Country
2015	Unified Ethical Frame for Big Data Analysis: IAF Big Data Ethics Initiative, Part A	The Information Accountability Foundation (2015)	Interprofessional group/ association	USA
2016	AI Now 2016 Report	AI Now Institute (2016)	Interprofessional group/ association	USA
	Digital decisions.	Center for Democracy & Technology	Interprofessional group/ association	USA
	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version for Public Discussion)	The IEEE Global Initiative on Ethics of Autonomus and Intelligent Systems	Interprofessional group/ association	International
	Position on Robotics and Artificial Intelligence	Green Digital Working Group (2016)	Interprofessional group/ association	Europe
	Report on the Future of Artificial Intelligence	Executive Office of the President National Science and Technology Council Committee on Technology (2016)	Government	USA
2017	AI Now 2017 Report	AI Now Institute at New York University	Interprofessional group/ association	USA
	ITI AI Policy Principles (2017)	Information Technology Industry Council	Interprofessional group/ association	Canada
	Machine Learning: The Power and Promise of Computers that Learn by Example	Royal Society (2017)	Interprofessional group/ association	UK
	Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society: Managing the Fourth Industrial Revolution	Government of the Republic of Korea	Government	Korea
	Report of COMEST on Robotics Ethics	COMEST/UNESCO (2017)	Inter-government	International
	Report on Artificial Intelligence and Human Society: Unofficial Translation	Ministry of State for Science and Technology Policy (2017)	Interprofessional group/ association	Japan
	Statement on Algorithmic Transparency and Accountability	Association for Computing Machinery (ACM 2017)	Interprofessional group/ association	USA
	The Asilomar AI Principles	Future of Life Institute (2017)	Interprofessional group/ association	International
	The Japanese Society for Artificial Intelligence Ethical Guidelines	Japanese Society for Artificial Intelligence (2017)	Interprofessional group/ association	Japan
	Top 10 Principles for Ethical AI	UNI Global (2017)	Interprofessional group/ association	International
	van Est, and Gerritsen (2017). Human Rights in the Robot Age: Challenges Arising from the Use of Robotics, Artificial Intelligence, and Virtual and Augmented Reality	Rathenau Institute	Inter-government	Europe

(*Continued*)

Table 1. (Continued).

Year	Title	Source	Type of document	Geographical scope/Country
2018	AI in the UK: Ready, Willing and Able 183	House of Lords (2018)	Government	UK
	AI Now 2018 Report	AI Now Institute	Interprofessional group/ association	USA
	AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations (2018)	AI4People – The first multi-stakeholder forum bringing together all actors interested in shaping the social impact of new applications of AI	Interprofessional group/ association	UK
	Artificial Intelligence at Google	Google (2018)	Private firm	USA
	Avila et al. (2018). Artificial Intelligence: Open Questions about Gender Inclusion	World Wide Web Foundation	Interprofessional group/ association	International
	Business Ethics and Artificial Intelligence	Institute of Business Ethics (2018)	Interprofessional group/ association	UK
	Charlevoix Common Vision for the Future of Artificial Intelligence	Leaders of the G7 (2018)	Inter-government	International
	Dutch Artificial Intelligence Manifesto	Special Interest Group on Artificial Intelligence	Interprofessional group/ association	Netherlands
	Everyday Ethics for Artificial Intelligence	IBM (2018)	Private firm	USA
	Montréal Declaration for Responsible Development of Artificial Intelligence (2018)	Scholars	Interprofessional group/ association	Canada
	Partnership on AI (2018)	Partnership on AI	Interprofessional group/ association	USA
	Science, law and society (SLS) initiative.	The Future Society (2018)	Interprofessional group/ association	International
	The Toronto declaration: protecting the right to equality and non-discrimination in machine learning systems (2018)	Human Rights Watch	Inter-government	Canada
	Villani, C. For a Meaningful Artificial Intelligence: Toward a French and European Strategy	AI for Humanity	Interprofessional group/ association	Europe
	White Paper: How to Prevent Discriminatory Outcomes in Machine Learning	World Economic Forum (2018)	Inter-government	International
2019	AI Now 2019 Report	AI Now Institute at New York University	Interprofessional group/ association	USA
	Beijing AI Principles	Beijing Academy of Artificial Intelligence (2019)	Government	China
	Dawson, D. et al. (2019). Artificial Intelligence: Australia's Ethics Framework	Australian Government	Government	Australia
	DeepMind Ethics & Society Principles	DeepMind (2019)	Private firm	UK
	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition) (2019)	The IEEE Global Initiative on Ethics of Autonomus and Intelligent Systems	Interprofessional group/ association	International

(*Continued*)

Table 1. (Continued).

Year	Title	Source	Type of document	Geographical scope/Country
	Initial code of conduct for data-driven health and care technology	GOV.UK (2019)	Government	UK
	Microsoft AI principles	Microsoft Corporation (2019)	Private firm	USA
	OECD Recommendation of the Council on Artificial Intelligence	Organization for Economic Cooperation and Development	Inter-government	International
	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (2019)	Scholars – Fairness, Accountability, and Transparency in Machine Learning	Interprofessional group/ association	International
	Responsible AI and robotics: an ethical framework	Accenture (2019)	Private firm	UK
	The responsible AI framework	PwC	Private firm	UK
2020	Livre blanc sur l'intelligence artificielle. Une approche européenne axée sur l'excellence et la confiance.	Commission Européenne	Inter-government	Europe
	Trustworthy AI framework	Deloitte AI Institute (2020)	Private firm	USA
2021	Recommandation sur l'éthique de l'intelligence artificielle	UNESCO (2021)	Inter-government	International

categories were discarded. Regarding DEI practical recommendations, we first decided to split recommendations for the implementation of DEI principles in two main categories targeting the objects of "AIS development process" and "AIS design organizations." The first deductive step allowed to identify within the main categories, respectively, 10 and 9 categories; the following inductive step led to discard one category in the "AIS design organizations," namely, "Penalties for noncompliance with DEI principles" since no entry matched it, resulting in, respectively, 10 and 8 categories (Table 2).

The content extraction, categorization, and analysis were performed manually and independently by two research graduate students in an AI educational program. Each assessed the content and extracted relevant parts within selected sources and classified them among categories. Extraction of contents consisted of full sentences or paragraphs to enhance comprehensive analysis in the next steps. To assess the degree of agreement between the two students after this first round, we estimated 3 kappa coefficients. One for the principles, a second for practical recommendations regarding AIS development process, and a third for practical recommendations regarding AIS design organizations. Results show an almost perfect agreement for the principles (k = 0.85; $p < .05$) and the AIS development process (k = 0.84; $p < .05$), and a substantial agreement for the regarding AIS design organizations (k = 0.80; $p < .05$) (Landis and Koch 1977). After this first round of analysis, they performed a second one to compare preliminary results and identify differences in extraction and/or categorization. A third researcher, specializing in DEI management in AI, inspected the whole results and assessed the consistency and

Table 2. Categories (deductive, deleted, and inductively added) and extracted contents.

High categories	Categories		Number of categories	Number of extracted contents	Part of extracted contents in the total/*subtotal*
EDI principles	EDI principles		14	361	38%
	Deductive	*Equity/Fairness*		*65*	*18%*
	Deductive	*Nondiscrimination*		*37*	*10%*
	Deductive	*Diversity in learning datasets/ Representativeness*		*27*	*7%*
	Deductive	*Human decisions made*		*28*	*8%*
	Deductive	*Inclusion*		*28*	*8%*
	Deductive	*Correction of existing inequalities in society/ positive action/ affirmative action*		*31*	*9%*
	Deductive	*Transparency*		*37*	*10%*
	Inductive added	*Human dignity*		*22*	*6%*
	Deductive	*Social justice*		*18*	*5%*
	Deductive	*Diversity in the AI sector/ Representativeness*		*22*	*6%*
	Deductive	*Accessibility/digital divide reduction*		*18*	*5%*
	Deductive	*DEI regulation conformity*		*15*	*4%*
	Inductive added	*Remedies for discrimination*		*9*	*2%*
	Inductive added	*Civil society interaction/ inclusion*		*4*	*1%*
EDI practical recommendations for implementing EDI principles	in the AIS development process		10	372	39%
	Deductive	*Verification and validation of AIS results*		*57*	*15%*
	Deductive	*AIS learning sources and dataset*		*50*	*13%*
	Deductive	*Transparency of AIS decisions*		*65*	*17%*
	Deductive	*Explicability of AIS decisions*		*40*	*11%*
	Deductive	*Functional design of AIS*		*46*	*12%*
	Deductive	*Technical design of AIS*		*25*	*7%*
	Deductive	*Correction of biased AIS results*		*24*	*6%*
	Deductive	*Mechanisms of redress for users/beneficiaries*		*30*	*8%*
	Deductive	*DEI Technical documentation for AIS design*		*25*	*7%*
	Deductive	*AIS DEI certification/labeling*		*10*	*3%*
	in AIS Design Organizations		8	216	23%
	Deductive	*Accountability/responsibility for AIS*		*45*	*21%*
	Deductive	*Team diversity in AIS companies*		*40*	*19%*
	Deductive	*DEI training/awareness*		*40*	*19%*
	Deductive	*Commitment to DEI principles/codes*		*33*	*15%*
	Deductive	*Ethics/DEI function implementation*		*18*	*8%*
	Deductive	*Multi-stakeholder inclusion for DEI governance*		*21*	*10%*
	Deductive	*Objectives/incentives for compliance with DEI principles*		*12*	*6%*
	Deductive	*Internal communication for compliance with DEI principles*		*7*	*3%*
	Deductive deleted	*Penalties for noncompliance with DEI principles*		*0*	*0%*
Total			32	949	100%

categorization. The final extraction process resulted in 361 entries for DEI principles and 588 different entries for DEI practical recommendations.

Results

Targeted Groups in DEI Principles in AI

DEI principles in AI always target certain groups to benefit from the erected principles, such as social justice or fairness. The three main groups we identified were female ($N = 34$), disabled or health vulnerable people ($N = 34$) and racial or ethnic minorities ($N = 32$). These groups mainly meet the identified groups for "affirmative action" or "employment equity" policies in western countries (Jain, Sloane, and Horwitz 2003). These proactive policies seek to achieve equality of treatment and relevant group representation within education and the workplace. It was also found that 61% of guidelines point to people with diverse sexual orientation ($N = 28$), almost half target natives, elderly people and religion or beliefs ($N = 22$), whereas youth is mentioned by 26% ($N = 12$), immigrants by 24% ($N = 11$), people far from the digital world (i.e., with low technological access and/or digital skills) by 22% ($N = 10$), and economically/socially disadvantaged people by 11% ($N = 5$). We can note that no source explicitly targets illiterate people.

DEI Principles in AI Ethical Guidelines

Most guidelines assert the importance of DEI in AI. As an example, in The Toronto Declaration, the Human Rights Watch (2018, 6), states that "*This Declaration underlines that inclusion, diversity and equity are key components of protecting and upholding the right to equality and nondiscrimination. All must be considered in the development and deployment of machine learning systems in order to prevent discrimination, particularly against marginalized groups.*" The content analysis allowed for the identification of 14 categories of principles related to DEI (Table 3). In their review of AI ethics principles, Hagendorff (2020) and Jobin, Ienca, and Vayena (2019) revealed that about 80% of their sources specify the notion of fairness. When focusing on DEI principles erected, our results tend to confirm these findings with 76% of sources mentioning "Equity/Fairness," by far one of the most cited notions, followed by "nondiscrimination" for 57%. These both DEI principles are related to the regulation of DEI in most western countries (Klarsfeld et al. 2014), so DEI principles in AI primarily remind actors of the DEI legal obligations already in place. A greater diversity in datasets used for developing or training AISs should provide a greater representativeness for half of the guidelines, thus delivering a more AI-specific principle and relying on one of the most often cited charges against discriminatory AI (Howard and Borenstein 2018). Echoing the important debate in AI about the accountability of decisions made and supported by AISs on individuals, especially when they can impact their personal life, health, safety or professional life, 46% of sources recommend that humans should always be responsible for final decisions made. Moreover, 41% of guidelines state that AISs should be sources of inclusion and should tend to correct existing inequalities in societies, playing a proactive role of improving the social environment in which the systems are developed or used and not only avoiding the reproduction of inequalities. Also, 41% of guidelines insist on the need for transparency of AI, what is "*about efforts to identify, prevent and mitigate against discrimination in AI systems*" (Toronto Declaration 2018, 12). The respect of human dignity and social justice (referred, respectively, to by 35% and 33% of guidelines) are also pointed out, referring to a consideration of all individuals without distinction.

One-third of the guidelines acknowledge the lack of diversity in the AI community (West, Whittaker, and Crawford 2019) and call for increasing the share of women and minorities in AI research and industries. This recommendation argues for the recognition that a more diverse representation in the AI sector would be beneficial for developing AISs better equipped to face DEI challenges, drawing from DEI literature demonstrating increased performance and better decision-making in teams with more diversity (Cox and Blake 1991; Sacco and Schmitt 2005). Also, 26% percent of guidelines consider the importance of addressing the digital divide, which entails that all have access to AISs developed, with adequate literacy

Table 3. DEI principles in AI (out of N of guidelines = 46) and associated DEI management paradigms.

	Number of sources	Distribution in all sources	Fair representation and treatment and non-discrimination	Valuing diversity (synergy)
Equity/Fairness	35	76%	x	
Nondiscrimination	26	57%	x	
Diversity in learning datasets/ Representativeness	23	50%	x	
Human decisions made	21	46%	x	
Inclusion	19	41%		x
Correction of existing inequalities in society/positive action/affirmative action	19	41%	x	
Transparency	19	41%	x	
Human dignity	16	35%	x	
Social justice	15	33%	x	
Diversity in the AI sector/ Representativeness	15	33%	x	
Accessibility/digital divide reduction	12	26%	x	
DEI regulation conformity	10	22%	x	
Remedies for discrimination	6	13%	x	
Civil society interaction/inclusion	4	9%		x
Total			12 (86%)	2 (14%)

and understanding, fighting against barriers to access and empowering beneficiaries. Finally, only 13% of guidelines recommend formal remedies for discrimination due to AI.

We then categorized DEI principles in AI alongside the DEI management paradigms (Table 3). Results show that 12 out of 14 principles fall under the paradigm of "fair representation and treatment and nondiscrimination." This suggests that AI guidelines most of all encourage compliance with fairness and nondiscrimination, to avoid creating, replicating or even amplifying existing unfair processes or results. Only two principles, "Inclusion" and "Civil society interaction/inclusion," were considered as echoing a perspective of valuing diversity to fuel and improve the AI development and use.

DEI Practices in AI Ethical Guidelines

The AIS Development Process

The content analysis allowed us to identify 10 categories of practical recommendations for implementing DEI principles in the AIS development process (Table 4). The most cited recommendation is to set mechanisms to monitor "verification and validation of AIS results" with regard to potential discrimination resulting of the AIS, including post-implementation (referred to by 63% of sources). "Functional design" and "technical design," which consider DEI-oriented practices from the beginning of the design phase of development, are recommended by only, respectively, 48% and 39% of AI guidelines. Yet, one would expect development practices to incorporate DEI principles at the design stage, rather than trying to correct deviance in the results obtained post-implementation. This can be explained by the fact that many AISs have already been developed and used by organizations, without any consideration for DEI issues at their early design stages, calling for vigilance on the outcomes produced by already running systems. This matters because all the more research suggests that discriminatory AISs could not easily be fixed post deployment: once a bias enters the algorithm, it becomes challenging to identify and eradicate it (Howard and Borenstein 2018). This was the case for the selection AIS deployed by Amazon in 2014, which discriminated against female applicants. The system could not be fixed and had to be removed from the selection process (Dastin 2018). Practices for ensuring more diverse and representative "AIS learning sources and dataset" is the second most recommended practice (by 57% of guidelines), in line with the principle mentioned by half of DEI guidelines in AI. Such practices recommend addressing bias in training data, assessing risks of datasets with historic or systemic bias or breaching discrimination laws, and ensuring datasets do not perpetuate social prejudices.

More than half of AI guidelines provide practical recommendations to enhance transparency (54%) and explicability (52%) of AISs decisions regarding DEI stakes.

This includes reliance on open data, environment and systems, and measures enabling external review and monitoring. Namely, the UNI Global (2017, 6) warns that “*clarity cannot be obfuscated by complexity*,” inviting to allow scrutiny of the system’s processes by independent entities. Explicability refers to a right to explanation for actions or decision made by AIS, avoiding to hide behind the AIS “black box” rationale often put forward. TheDutch Artificial Intelligence (2018, 5) invites to “*develop new algorithms, which (1) by design can explain their rationale, (2) do so in an intuitive, human-understandable manner, and (3) explain why their underlying mechanisms produced the AI’s behavior*.” Some guidelines (35%) suggest recourse mechanisms for users or beneficiaries whose rights are violated or abused through the use of AISs. These recourse mechanisms would allow to appeal an AIS’s automated decision and allow for a new, non-automated decision, the opportunity to refer to an AI ombudsperson to ensure the auditing of allegedly unfair or inequitable results of AI, or the introduction of specialized insurance to ensure user protection. About 30% of guidelines call to produce technical documentation for and by AIS designers to support the practical application of DEI principles, particularly about avoiding bias that could lead to discrimination. The Australian Government (2019, 8) even proposes “*the provision of educational guides, training programs [. . .] to help implement ethical standards in AI use and development*..” Finally, only 15% of guidelines encourage for a “DEI certification/labeling of AIS” by external and independent entities which would integrate an evaluation of AIS for their inclusiveness or nondiscrimination performance. It is most often suggested as a voluntary approach with the objective of strengthening user confidence. We note that no AI guideline call for strong regulation regarding DEI in the AI development process, rather favoring a “soft” law approach through labels or ISO-type certifications.

AIS Design Organizations

Regarding practical recommendations for implementing DEI principles in AIS design organizations, eight categories of practices were identified (Table 5). More than half of the guidelines (57%) call organizations for

Table 4. Recommended practices for implementing DEI principles in the AIS development process (out of N of guidelines = 46).

	Number of sources	Percentage of all sources
Verification and validation of AIS results	29	63%
AIS learning sources and dataset	26	57%
Transparency of AIS decisions	25	54%
Explicability of AIS decisions	24	52%
Functional design of AIS	22	48%
Technical design of AIS	18	39%
Correction of biased AIS results	18	39%
Mechanisms of redress for users/beneficiaries	16	35%
DEI Technical documentation for AIS design	14	30%
AIS DEI certification/labeling	7	15%

integrating, analyzing and holding the responsibility of AIS decisions and actions, in terms of legacy, social and/or technical accountability. Which means that AIS design organizations are responsible for identifying which DEI legislation applies to them. This requires the development of procedures, tools and methods to audit the systems and evaluate their conformity to legal and ethical frameworks (Villani 2018). This also reminds us that until now, AISs themselves are not responsible parties under the law, even if some call for reforming legal systems to grant rights and responsibilities to AI and self-learning machines (Government of the Republic of Korea 2017). Some have argued that existing national or international antidiscrimination laws provide sufficient guidance on how to regulate AI and their impacts on DEI. While others have highlighted the need to establish specific AI standards and regulatory bodies for addressing issues of discrimination, bias, and unfairness at the different phases of the life cycle, considering the accelerating pace of AI in various spheres (Wallach and Marchant 2018). Additionally, four suggested practices relate directly to well-established actions when managing DEI within organizations: diversifying AI teams within the organization (50%), providing DEI training to raise awareness of designers, developers, and trainers about DEI issues and risks of AISs (50%), committing to a DEI charter or code for AI and/or developing such internal policy (35%), putting in place a dedicated ethics/DEI function (22%). Some directives (IEEE 2016) underline that even if having an ethics/DEI function is a good practice, responsibility should not lie solely with them. Instead, all team members should act responsibly throughout the AI design process. In addition, 20% of guidelines encourage the development of an inclusive governance for DEI stakes implying multiple stakeholders when designing, but also testing and improving AISs. Moreover, 13% of sources call for clear objectives or incentives, including financial ones, for compliance with DEI principles, to foster diversity within the organization or to meet DEI principles in AIS development. Some also recommend that conscientious objectors and workers raising DEI concerns should be protected (Whittaker et al. 2018). And 9% of the guidelines highlight the role played by internal communication for complying with DEI principles.

Finally, we categorized the DEI practices along the three components of a positive DEI climate, which reflect the DEI system and practices perceived by an organization's incumbents (Cachat-rosset, Carillo, and Klarsfeld 2019). Three out of eight recommended organizational practices meet the intentionality dimension for managing diversity, and four meet the programmatic one (Table 5). Only one, directly addresses the praxis dimension, that is the day-to-day attitudes and behaviors of individuals and teams along DEI principles. Furthermore, this practice is only mentioned by 13% of the guidelines.

Table 5. Recommended practices for implementing DEI principles in AIS design organizations (N of guidelines = 46) and associated DEI climate dimensions.

Recommended practices for implementing DEI principles in the AIS development process	Number of sources	Percentage of all sources	DEI climate dimensions		
			Intentionality	Programmatic	Praxis
Accountability/responsibility for AIS	26	57%		x	
Team diversity in AIS companies	23	50%		x	
DEI training/awareness	20	43%		x	
Commitment to DEI principles/codes	16	35%	x		
Ethics/DEI function implementation	10	22%		x	
Multi-stakeholder inclusion for DEI governance	9	20%	x		
Objectives/incentives for compliance with DEI principles	6	13%			x
Internal communication for compliance with DEI principles	4	9%	x		
Total			3 (37.5%)	4 (50%)	1 (12.5%)

Discussion

Place and Role of DEI Principles in AI (R1a)

Our literature review of DEI principles in AI found no document, code or charter dedicated to the DEI topic in AI, within our research criteria of recent, reliable sources with a large scope of influence. The DEI principles that we found were always included in more global ethical principles for AI. This contrasts with other ethical principles in AI such as data privacy and associated remedies, which have benefited from dedicated guidelines or regulations, such as the General Data Protection Regulation (GDPR) in Europe, or the Privacy Act in Canada, that protect the privacy of individuals with respect to personal information and provide them with rights regarding access, use or correction of this information. To date, the principles of DEI in AI have not received the same level of attention.

In this context, DEI principles for the development and use of fair, equitable and nondiscriminatory AI represent a step toward a "soft" framework, based on voluntary compliance (Campolo et al. 2017). This approach follows the stakeholder theory (Freeman 1984), whereby companies must meet the needs of their stakeholders – i.e., all the people affected by the decisions they make – to survive and then make a profit. Such a theory would provide an effective means toward adoption of DEI principles by AI industries, when stakeholders are fully empowered to exert pressure. 1997) characterized three cumulative attributes for stakeholders to have a proper influence: (1) power of influence and constraints, based on the resources they control; (2) legitimacy – a general perception or assumption that an entity's activities are desirable or appropriate to some socially constructed system of norms, values, and beliefs; and (3) urgency – when the actors feel that their demand is pressing or important. The strong affirmation of the need for adopting DEI principles in AI lends credence to the legitimacy of the demands of individuals who benefit from AIS-based decisions, and the velocity with which many national and international bodies have decided to establish these principles point to the urgency to these demands. But the power of influence of stakeholders is questionable, as public understanding of AI technologies is often limited (Curtis, Gillespie, and Lockey 2022; Selwyn and Gallo Cordoba 2021). Some scholars even revealed that the fairness discourse in AI is largely co-produced by tech companies and associations to avoid further fair regulation and to preserve their interests (Ochigame 2019; Weinberg 2022). On the other hand, about 19 out of 46 guidelines for AI (41%) advocate that AISs should not only be exempt of unfair or discriminatory processes and results but also that they should embrace a proactive role in redressing the unbalanced standing of marginalized groups relative to majority groups. In this vein, AISs are expected to be used as innovative tools for detecting and alerting of existing discrimination in processes and reduce inequalities in societies.

Limited Paradigms and Contexts-Sensitivity of DEI Principles and Targets (R1b)

Our results show that 86% of AI guidelines meet the "fair representation and treatment and nondiscrimination" paradigm for managing DEI (Dwertmann, Nishii, and van Knippenberg 2016; Thomas and Ely 1996). The primary approach to DEI principles in AI is therefore to respect the norms and values of equity and social justice, in addition to complying with anti-discrimination laws. Albeit valuable, this is a reactive approach consisting in avoiding liability for noncompliance with legal and/or socially acceptable frameworks, rather than a proactive consideration of the benefits of diversity in AI (i.e. the synergy perspective). The studied guidelines come from private firms or professional association for 70% of them, and (inter-)government for 30%. So, we suggest that future studies further investigate if DEI principles follow different paradigms and are given more compliant or proactive aim when coming from corporate or government public policies for AI. Moreover, we highlight that seven targeted groups are widely shared by the different guidelines in AI (i.e., Gender (74%), Disabled/Health vulnerable (74%), Racial/Ethnic minorities (70%), Sexual orientation (62%), Natives (49%), Age/Older people (49%) and Religion/Beliefs (49%)). This suggests that there is a widely shared approach of equity and social justice principles and of discriminated groups at stake in the AI guidelines. However, the DEI literature has emphasized the strong contingent nature of DEI understanding and management (Syed and Özbilgin 2009). Whatever the followed paradigm, DEI assumptions are highly influenced, on the one hand, by national and social cultures, discourses, history, toleration of discrimination, groups' domination and representations (Ng and Klarsfeld 2018), and on the other hand, by industries, organizational structures, DEI maturity or resources invested in DEI management (Djabi-Saïdani and Pérugien 2019). As such, DEI is a social construct regarding targeted groups and regarding social, business or inclusive paradigm at stake. Hanna et al. (2020, 2) also argue that AI fairness research does not sufficiently consider how the social group categories, operationalized in AI guidelines, are socially constructed, resulting in the "widespread use of racial categories as if they represent natural and objective differences between groups." Thus, we suggest to challenge the targeted groups in each context in which the AIS is used and that other categories of potentially discriminated groups, than the seven widely mentioned, could be more relevant to non-Western contexts (e.g., castes in India), and would deserve being represented too.

Regarding AI guidelines, the tone is largely set by Western countries in terms of DEI principles (65% come from North America, Europe and Australia), promulgated as a one-way approach with few considerations or adaptations to countries with different cultures. A limitation of this review is that we selected only sources written in English, which may explain the strong domination of Western countries in our results even if documents collected in non-Western areas (e.g., from Asia) were also available in English and were therefore included. Donaldson (1989), in discussing ethical norms, suggests that there is a coexistence of cross-cultural "hyper-norms" with culture- and region-specific norms, even as global international practices expand. So, if DEI principles in AI from Western countries are enacted as new worldwide shared standards without local amendments, these principles may be unsuitable for non-Western contexts and social environments (Kiemde and Kora 2022), which at best, risks inefficiency and inoperability, and at worst the introduction of new biases or discrimination. Hagendorff (2021) also emphasized that it may be too simplistic to formally encode the concepts of justice or fairness in AISs as fixed topics whereas they are relational and social constructs. We thus encourage future research to investigate the influence of Western domination in DEI principles for AI, and how specific DEI approaches should be considered regarding various cultural contexts, as well as to question whether it is appropriate to transfer AISs developed in a specific cultural context (i.e., Western) to another one (i.e., non-Western).

In addition, if 21% of AI guidelines target people remote from the digital world (i.e., with low technological access and/or digital skills) and 11% the economically/socially disadvantaged, the definition of these targets and the impact of AISs in more or less developed countries could be questioned in future research. If AI literacy begins to be a subject of research and recommendations (Long and Magerko 2020; Ng et al. 2021), such research and recommendations should also explore the technological distance and literacy from diverse economic and cultural contexts.

The Rather Technical Operationalization of DEI Principles in AI Practices (R2a)

Despite the promulgation of ethical principles in AI for greater fairness and equity (Jobin, Ienca, and Vayena 2019), statements remain general, very broad (Hagendorff 2020) and rarely address concrete implementation and results (Crawford et al. 2019). The analyzed documents provide some guidance for AI actors concerned with technical solutions (e.g., verification and validation processes, diverse learning sources and datasets), individuals' involvement (e.g., DEI awareness, objectives/incentives for compliance) and organizational actions (e.g., increased team diversity, organizational commitment to DEI, Ethics/DEI function). However, there is still a notable gap between the practices proposed in the literature and the challenges faced by AI designers in day-to-day operations (Holstein et al. 2019). Moreover, we note that 7 out of 10 practices devoted to the AIS development process are recommended by 40% and more of guidelines, showing a quite strong consensus on technical answers to DEI stakes when developing AISs. Whereas non-technical recommendations, i.e. those devoted to improving DEI in AIS Design Organizations themselves, do not meet the same level of consensus, with only three practices out of eight that are recommended by at least 40% of the guidelines. This result reveals that technical practices are more widely shared than organizational ones when it comes to enhancing DEI in AI. Whereas Schiff et al. (2020) highlighted that it is not only technical practices that must adapt for ethics in AI but also organizational practices and integrated teams with technical and non-technical profiles.

Many AI guidelines indeed suggest that ethical challenges are best addressed through technical and design expertise (Greene, Hoffmann, and Stark 2019). Namely, Jobin, Ienca, and Vayena (2019) propose actions in ethical guidelines to include principles of justice in AI, the first three actions being: technical solutions, such as standards or normative encoding, followed by transparency, and then technical tests and audits. Whereas adopting a more interdisciplinary and inclusive view, including a more diverse workforce and relevant stakeholders from the civil society, or undertaking systemic changes, only come last in the proposed actions (Jobin, Ienca, and Vayena 2019). Critical research and alternative perspectives are largely neglected (Hagendorff 2021). West, Whittaker, and Crawford (2019) argue that AI fairness issues must go beyond technical debiasing to include a broader social analysis of how AI is used in context, so that a more capacitive accounting of bias becomes possible. Alongside Mittelstadt (2019, 10), we invite actors to "*pursue [AI] ethics [and DEI] as a process, not technological solutionism.*"

In scrutinizing the 120 identified authors for the 46 selected guidelines, we found that 60.8% are male, 74.2% are white and 78.3% are nationals of Western countries (from North America, Europe or Australia). Further research should determine whether the predominance of white males among guideline authors is linked to the predominance of technical solutions. It should also investigate whether white male dominance is linked to considering DEI in AI as an isolated technological issue to be fixed rather than one embedded within its social context, such as previously pointed out for moral problems (Gilligan 1982) or ethical stakes (Hagendorff 2020). The lack of diversity among authors of ethical guidelines echoes the lack of diversity within the AI community. The AI Now report (West, Whittaker, and Crawford 2019) already pointed out that women made up only 15% of AI research staff at Facebook and 10% at Google, and that of the workforce at Google only 2.5% were black, and 4% at Facebook and Microsoft. This is not conducive to a pro-diversity climate in AIS design organizations (Cachat-rosset, Carillo, and Klarsfeld 2019 So, we encourage future research to investigate a gendered- and a minority-based approach of DEI operationalization in AI that would embrace a less technical view in favor of a more holistic, interdisciplinary, and challenging one.

The Limited Place of Attitudes and Behaviors Among Targets of DEI Recommended Practices (R2b)

Our results show that only one recommended practice for DEI operationalization in AI directly addresses the praxis dimension of a DEI climate (Objectives/ incentives for compliance), whereas all seven other practices for DEI operationalization in AI aim at promoting the intentionality or programmatic dimension in AI organizations. Not even one of the DEI guidelines included one of the categories first considered for practices, namely "Penalties for noncompliance with DEI principles," which could have helped to directly orient behaviors toward the operationalization of DEI. Whether intentions and pro-DEI programs may alone positively influence DEI behaviors of AI actors, or binding regulation is needed such as being discussed in the EU (Voss 2021) has not yet been directly studied and deserves further research efforts. More research is also needed to ascertain that practical DEI recommendations in AI guidelines are sufficient for enhancing DEI behaviors. In addition, the lack of diversity in AIS design organizations is not conducive to shaping a positive DEI climate.

Furthermore, literature finds that codes of conduct do not impact behaviors or decision-making in technologies (McNamara, Smith, and Murphy-Hill 2018). Even if ethical codes or policies improve awareness about ethical issues in some cases, they are not sufficient for ensuring fairness (Van den Bergh and Deschoolmeester 2010). To better encompass marginal conditions to influence the adoption of DEI behaviors in AI, we also suggest considering Elango's (2010) comprehensive model for ethical intention and behaviors. It posits that congruence between individual ethics and organizational ethics is necessary to foster intention to adopt ethical behaviors. In contextualizing this model, we assert that both individual awareness of DEI stakes in AI and organizational support for DEI practices would be required for influencing DEI intention and behaviors. Unfortunately, DEI awareness is not very present in AI education and organizations. As an example in Quebec, a prime location for AI according to the Tortoise Global AI Index (Benessaieh 2022), an overview of educational programs addressing AI revealed that fewer than 20% provide clear ethics content, and none dedicated DEI content. This confirms that there is room for improving AI actors DEI awareness. This is even more prevalent in less developed and mature countries with respect to the development of AI such as in Africa (Kiemde and Kora 2022). Therefore, future research and practice should better address individual awareness of DEI stakes in AI and its impact on DEI behaviors. This could complement organizational actions in support of DEI, and help set up a DEI climate conducive to DEI-friendly behaviors when designing AISs, and more generally within AIS design organizations.

Conclusion

In this paper, a semi-systematic literature review of the guidelines regarding DEI in AI was performed using extant DEI literature. We unpacked their 14 underlying principles and their 18 recommended practices. Based on the analysis of these principles and recommended practices, we conclude that DEI principles put forward in the guidelines mainly aim at encouraging fairness, justice and non-discrimination in a limited compliance approach, ignoring other possibilities opened up by the DEI literature. We then identified that technical practices are favored to remedy DEI stakes in AI, over a more encompassing social and relational approach. Finally, we conclude that recommended practices for implementing DEI principles in AISs neglect actions aimed at directly influencing AI actors' behaviors and individual awareness of DEI risks and best practices.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Gaelle Cachat-Rosset http://orcid.org/0000-0001-7050-1677

References

Accenture. 2019. *Responsible AI and robotics: An ethical framework*. Accenture, UK.

ACM. 2017. *Statement on algorithmic transparency and accountability*. Washington, DC: Association for Computing Machinery US Public Policy Council.

Avila, R., A. Brandusescu, J. O. Freuler, and D. Takur. 2018. *Artifcial intelligence: Open questions about gender inclusion*. Argentina: World Wide Web Foundation.

Beijing Academy of Artificial Intelligence. 2019. Beijing AI Principles. *Beijing Academy of Artificial Intelligence (BAAI)*, Beijing: Beijing Academy of Artifcial Intelligence.

Benessaieh, K. 2022. Le Québec se classe 7e au monde. *La Presse*, 29(2): March 9th.

Bolukbasi, T., K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems* 29:4349–57.

Bowen, D. E., and C. Ostroff. 2004. Understanding HRM–firm performance linkages: The role of the "strength" of the HRM system. *Academy of Management Review* 29 (2):203–21.

Boxenbaum, E. 2006. Lost in translation, the making of Danish diversity management. *The American Behavioral Scientist* 49 (7):939–48. doi:10.1177/0002764205285173.

Bradley, T. 2017 (July, 31). Facebook AI Creates Its Own Language in Creepy Preview of Our Potential Future. Forbes. https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/?sh=4367685f292c

Cachat-rosset, G., K. A. Carillo, and A. Klarsfeld. 2019. Reconstructing the concept of diversity climate–a critical review of its definition, dimensions, and operationalization. *European Management Review* 16 (4):863–85. doi:10.1111/emre.12133.

Cachat-rosset, G., K. A. Carillo, and A. Klarsfeld. 2021. Exploring the impact of diversity climate on individual work role performance: A novel approach. *European Management Review* 19 (2):248–62. doi:10.1111/emre.12483.

Campolo, A., M. Sanflippo, M. Whittaker, and K. Crawford 2017. AI Now Report 2017. AI Now Institute, New York.

Chung, Y., H. Liao, S. E. Jackson, M. Subramony, S. Colakoglu, and Y. Jiang. 2015. Cracking but not breaking: Joint effects of faultline strength and diversity climate on loyal behavior. *Academy of Management Journal* 58 (5):1495–515. doi:10.5465/amj.2011.0829.

COMEST/UNESCO. 2017. Report of COMEST on robotics ethics. COMEST/UNESCO SHS/YES/COMEST-10/17/2 REV, Paris.

Cox, T. H., and S. Blake. 1991. Managing cultural diversity: Implications for organizational competitiveness. *Academy of Management Perspectives* 5 (3):45–56. doi:10.5465/ame.1991.4274465.

Crawford, K., Dobbe R, Dryer T, Fried G, Green B, Kaziunas E, Kak A, Mathur V, McElroy E, Nill Sánchez A,et al. 2019. AI Now Report 2019. AI Now Institute, New York.

Crawford, K., and M. Whittaker 2016. The AI now report: The social and economic implications of artificial intelligence technologies. AI Now Institute, New York.

Curtis, C., N. Gillespie, and S. Lockey. 2022. AI-deploying organizations are key to addressing 'perfect storm' of AI risks. *AI and Ethics* 1–9. doi:10.1007/s43681-022-00163-7.

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women - reuters. Reuters :5–9. Accessed August 9, 2022. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dawson, D., E. Schleiger, J. Horton, J. McLaughlin, C. Robinson, G. Quezada, J. Scowcroft, and S. Hajkowicz. 2019. *Artificial intelligence: Australia's ethics framework*. Data61 CSIRO, Australia.

DeepMind. 2019. *Ethics and society principles*. London: DeepMind.

Deloitte. 2020. *Trustworthy AI*. USA: The Deloitte AI Institute.

Diakopoulos, N., Friedler S, Arenas M, Barocas S, Hay M, Howe B, Jagadish HV, Unsworth K, Sahuguet A, Venkatasubramanian S,et al. 2019. Principles for accountable algorithms and a social impact statement for algorithms. *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*. https://www.fatml.org/resources/principles-for-accountable-algorithms

Djabi-saïdani, A., and S. Perugien. 2019. The shaping of diversity management in France: An institutional change analysis. *European Management Review* 17 (1):229–46. doi:10.1111/emre.12343.

Donaldson, T. 1989. *The Ethics of International Business*. OxfordNew York.

Dutch Artificial Intelligence. 2018. *Dutch artificial intelligence manifesto*. The Netherlands: Special Interest Group on Artificial Intelligence.

Dwertmann, D. J. G., L. H. Nishii, and D. van Knippenberg. 2016. Disentangling the fairness and discrimination and synergy perspectives on diversity climate: Moving the field forward. *Journal of Management* 42 (5):1136–68. doi:10.1177/0149206316630380.

Elango, B., K. Paul, S. K. Kundu, and S. K. Paudel. 2010. Organizational ethics, individual ethics, and ethical intentions in international decision-making. *Journal of Business Ethics* 97:543–61.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profle, police, and punish the poor*. New York: St. Marting's Press.

European Commission. 2020. The Assessment List for Trustworthy Artificial Intelligence for self assessment. In *Directorate-General for Communications Networks*. Content and Technology, Publications Office.

Executive Office of the President National Science and Technology Council Committee on Technology. 2016. Report on the future of artificial intelligence, USA.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al. 2018. Ai4people—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*. 28 (4):589–707. doi:10.1007/s11023-018-9482-5.

Freeman, R. E. 1984. Strategic management: A stakeholder theory. *Journal of Management Studies* 39 (1):1–21.

Future of Life Institute. 2017. *Asilomar AI principles*. California: Future of Life Institute.

The Future Society. 2018. *The future society, law & society initiative, principles for the governance of AI*. Policy Research, The Law & Society Initiative.

Garcia-Gathright, J., and A. Springer 2018. Assessing and addressing algorithmic bias – but before we get there. In 2018 AAAI Spring Symposium Series, California.

Gilligan, C. 1982. *In a different voice: Psychological theory and women's development*. Cambridge: Harvard University Press.

Google. 2018. *Artificial intelligence at google: Our principles*. Google AI.

Government of the Republic of Korea. 2017. *Mid- to long-term master plan in preparation for the intelligent information society: Managing the fourth industrial revolution*. Korea: Government of the Republic of Korea Interdepartmental Exercise.

GOV.UK. 2019. *Initial code of conduct for data-driven health and care technology*. United Kingdom: GOV.UK.

Green Digital Working Group. 2016. *Position on robotics and artificial intelligence*. The Green Digital Working Group, Europe.

Greene, D., A. L. Hoffmann, and L. Stark 2019. *Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning*. Proceedings of the 52nd Hawaii International Conference on System Sciences, Hawaii.

Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines* 30 (1):99–120. doi:10.1007/s11023-020-09517-8.

Hagendorff, T. 2021. Blind spots in AI ethics. *AI Ethics* (4):1–17. doi:10.1007/s43681-021-00122-8.

Hanna, A., E. Denton, A. Smart, and J. Smith-Loud 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain.

Holstein, K., J. Vaughan, H. Daumé Iii, M. Dudik, and H. Wallach 2019. Improving fairness in machine learning systems: What do industry practitioners need? In 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK.

House of Lords. 2018. AI in the UK: Ready, willing and able? House of Lords Select Committee on Artificial Intelligence, HL Paper 100, London.

Howard, A., and J. Borenstein. 2018. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and Engineering Ethics* 24 (5):1521–36. doi:10.1007/s11948-017-9975-2.

IBM. 2018. *Everyday ethics for artificial intelligence, IBM design for AI*. USA.

IEEE. 2016. *Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems, version 1*. IEEE Advancing Technology for Humanity. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v1.pdf

IEEE. 2019. *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*, First Edition ed, IEEE Advancing Technology for Humanity. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9398613

Information Accountability Foundation. 2015. *Unified ethical frame for big data analysis: IAF big data ethics initiative*. Part A.

Institute of Business Ethics. 2018. *Business ethics and artificial intelligence*. London: Institute of Business Ethics.

ITI. 2017. *AI policy principles*. Information Technology Industry Council. https://www.itic.org/public-policy/ITIAIPolicyPrinciplesFINAL.pdf

Jain, H. C., P. J. Sloane, and F. M. Horwitz. 2003. *Employment equity and affirmative action: An international comparison*. New-York: ME Sharpe.

Japanese Society for Artificial Intelligence. 2017. Ethical Guidelines. In *The Japanese Society for Artificial Intelligence*, Japan.

Jobin, A., M. Ienca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1 (9):389–99. doi:10.1038/s42256-019-0088-2.

Kiemde, S. M. A., and A. D. Kora. 2022. Towards an ethics of AI in Africa: Rule of education. *AI and Ethics* 2 (1):35–40. doi:10.1007/s43681-021-00106-8.

Klarsfeld, A., L. A. E. Booysen, E. Ng, I. Roper, and A. Tatli. 2014. *Perspectives from 16 countries on diversity and equal treatment at work: An overview and transverse questions country perspectives on diversity and equal treatment*. Cheltenham: Edward Elgar Publishing.

Klarsfeld, A., L. Knappert, A. Kornau, F. W. Ngunjiri, and B. Sieben. 2019. Diversity in under-researched countries: New empirical fields challenging old theories? *Equality, Diversity and Inclusion: An International Journal* (7):694–704. doi:10.1108/EDI-03-2019-0110.

Kulik, C. T. 2014. Working below and above the line: The research-practice gap in diversity management. *Human Resource Management Journal* 24 (2):129–44. doi:10.1111/1748-8583.12038.

Landis, J., and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 (1):159–74. doi:10.2307/2529310.

Larsson, S., and F. Heintz. 2020. Transparency in artificial intelligence. *Internet Policy Review* 9 (2). doi:10.14763/2020.2.1469.

Leaders of the G7. 2018. *Common vision for the future of artificial intelligence, G7*. Charlevoix.

Long, D., and B. Magerko 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, Honolulu HI USA, 1–16.

Lorbiecki, A., and G. Jack. 2000. Critical turns in the evolution of diversity management. *British Journal of Management*, Special Issue, 17-31. s1. 10.1111/1467-8551.11.s1.3

McNamara, A., J. Smith, and E. Murphy-Hill 2018. Does ACM's code of ethics change ethical decision making in software development? In 26th ACM Joint ESE Conference and Symposium on the FSE:729–33.

Microsoft Corporation. 2019. *Microsoft AI principles*. USA: Microsoft Corporation.

Ministry of State for Science and Technology Policy. 2017. Report on Artificial Intelligence and Human Society: Unofficial Translation. In *Advisory Board on Artificial Intelligence and Human Society*, Japan.

Mitchell, R. K., B. R. Agle Et, and D. J. Wood. 1997. Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review* 22 (4):85386. doi:10.2307/259247.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1 (11):501–07. doi:10.1038/s42256-019-0114-4.

Moher, D., A. Liberati, J. Tetzlaf, D. G. Altman, and The PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* 6 (7):e1000097. doi:10.1371/journal.pmed.1000097.

Montréal Declaration. 2018. *Montréal declaration for a responsible development of artificial intelligence*. Montreal. https://www.montrealdeclaration-responsibleai.com/

Morley, J., L. Floridi, L. Kinsey, and A. Elhalal. 2019. A typology of ai ethics tools, methods and research to translate principles into practices. Retrieved 12 2022. *from*. https://aiforsocialgood.github.io/neurips2019/accepted/track2/pdfs/26_aisg_neurips2019.pdf

Ng, E. S., and A. Klarsfeld. 2018. Comparative and multi-country research in equality, diversity and inclusion. In *Handbook of research methods in diversity management, equality and inclusion at work*, 122–46. Edward Elgar Publishing.

Ng, D. T. K., J. K. L. Leung, K. W. S. Chu, and M. S. Qiao 2021. AI literacy: Definition, teaching, evaluation and ethical issues. In *Proceedings of the Association for Information Science and Technology*, Salt Lake City, UT, *58*(1): 504–09.

Nyholm, S., and J. Smids. 2016. The ethics of accident-algorithms for self-driving cars: An applied trolley problem?. *Ethical Theory and Moral Practice* 19 (5):1275–89.

Ochigame, R. (2019, December 20). The invention of 'ethical AI': How big tech manipulates academia to avoid regulation. *The Intercept*. https://theintercept.com/2019/12/20/mit-ethical-aiartificial-intelligence/

OECD. 2019. *Recommendation of the council on artificial intelligence* . OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449#:~:text=The%20OECD%20Council%20adopted%20the,on%2022%2D23%20May%202019.&text=The%20OECD%20Recommendation%20on%20AI,governments%20in%20their%20implementation%20efforts.

O'neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Oswick, C., and M. Noon. 2014. Discourses of diversity, equality and inclusion: Trenchant formulations or transient fashions? *British Journal of Management* 25 (1):23–39. doi:10.1111/j.1467-8551.2012.00830.x.

Partnership on AI. 2018. *Advancing positive outcomes for people and society*.

PwC. 2019. *The responsible AI framework*. PwC.

Royal Society. 2017. *Machine learning: The power and promise of computers that learn by example*. London: The Royal Society.

Sacco, J. M., and N. Schmitt. 2005. A dynamic multilevel model of demographic diversity and misfit effects. *The Journal of Applied Psychology* 90 (2):203–31. doi:10.1037/0021-9010.90.2.203.

Schiff, D., J. Biddle, J. Borenstein, and K. Laas 2020. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20*, 153–58. NewYork, NY, USA: Association for Computing Machinery.

Selbst, A. D., D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain.

Selwyn, N., and B. Gallo Cordoba. 2021. Australian public understandings of artificial intelligence. *AI & Society* 37 (4):1645–62. doi:10.1007/s00146-021-01268-z.

Singh, B., D. E. Winkel, and T. T. Selvarajan. 2013. Managing diversity at work: Does psychological safety hold the key to racial differences in employee performance? *Journal of Occupational & Organizational Psychology* 86 (2):242–63. doi:10.1111/joop.12015.

Snyder, H. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research* 104:333–39. doi:10.1016/j.jbusres.2019.07.039.

Syed, J., and M. Özbilgin. 2009. A relational framework for international transfer of diversity management practices. *The International Journal of Human Resource Management* 20 (12):2435–53. doi:10.1080/09585190903363755.

Thomas, D. A., and R. J. Ely. 1996. Making differences matter: A new paradigm for diversity management. *Harvard Business Review* 74 (5):79–90.

Toronto declaration. 2018. *The toronto declaration, protecting the right to equality and non-discrimination in machine learning systems*. Toronto: Human Rights Watch.

Tranfield, D., D. Denyer, and P. Smart. 2003. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management* 14 (3):207–22. doi:10.1111/1467-8551.00375.

UNESCO. 2021. *Recommendation on the ethics of artificial intelligence*. Paris: SHS/BIO/REC-AIETHICS/2021.

UNI Global Union. 2017. *Top 10 Principles for Ethical AI*. Switzerland: UNI Global Union.

Van den Bergh, J., and D. Deschoolmeester. 2010. Ethical decision making in ICT: Discussing the impact of an ethical code of conduct. *Communications of the IBIMA* 1–10. doi:10.5171/2010.127497.

van Est, R., and J. Gerritsen. 2017. *Human rights in the robot age: Challenges arising from the use of robotics, AI, and virtual and augmented reality*. The Netherlands: Rathenau Instituut.

Villani, C. 2018. *For a meaningful artificial intelligence: Towards a French and European strategy*. France: AI for Humanity.

Voss, G. 2021. AI act: The European union's proposed framework regulation for artificial intelligence governance. *Journal of Internet Law* 25 (4):1, 8–17. (2021).

Wallach, W., and G. E. Marchant 2018. An agile ethical/legal model for the international and national governance of AI and robotics. *Aies Conference on Artificial Intelligence, Ethics and Society*, New Orleans, USA.

Weinberg, L. 2022. Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ML fairness approaches. *The Journal of Artificial Intelligence Research* 74:75–109. doi:10.1613/jair.1.13196.

West, S. M., M. Whittaker, and K. Crawford. 2019. *Discriminating systems: Gender, race and power in AI*. AI Now Institute.

Whittaker, M., Crawford K, Dobbe R, Fried G, Kaziunas E, Mathur V, West SM, Richardson R, Schultz J, Schwartz O et al. 2018. *AI now report 2018*. New York: AI Now Institute.

World Economic Forum. 2018. How to Prevent Discriminatory Outcomes in Machine Learning. In *World Economic Forum Global Future Council on Human Rights 2016-18, REF*, 120318–case 00040065 . Switzerland.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *ArXiv: 170709457*.

Zou, J., and L. Schiebinger. 2018. Design AI so that it's fair. *Nature* 559 (7714):324–26. doi:10.1038/d41586-018-05707-8.

Zuboff, S. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: Public Affairs.