Research Article

Emerging Consensus on 'Ethical AI': Human Rights Critique of Stakeholder Guidelines

Sakiko Fukuda-Parr
The New School
Elizabeth Gibbons
FXB Center for Health and Human Rights, and Harvard T.H. Chan School of Public Health, and Harvard University

Abstract

Voluntary guidelines on 'ethical practices' have been the response by stakeholders to address the growing concern over harmful social consequences of artificial intelligence and digital technologies. Issued by dozens of actors from industry, government and professional associations, the guidelines are creating a consensus on core standards and principles for ethical design, development and deployment of artificial intelligence (AI). Using human rights principles (equality, participation and accountability) and attention to the right to privacy, this paper reviews 15 guidelines preselected to be strongest on human rights, and on global health. We find about half of these ground their guidelines in international human rights law and incorporate the key principles; even these could go further, especially in suggesting ways to operationalize them. Those that adopt the ethics framework are particularly weak in laying out standards for accountability, often focusing on 'transparency', and remaining silent on enforceability and participation which would effectively protect the social good. These guidelines mention human rights as a rhetorical device to obscure the absence of enforceable standards and accountability measures, and give their attention to the single right to privacy. These 'ethics' guidelines, disproportionately from corporations and other interest groups, are also weak on addressing inequalities and discrimination. We argue that voluntary guidelines are creating a set of de facto norms and re-interpretation of the term 'human rights' for what would be considered 'ethical' practice in the field. This exposes an urgent need for action by governments and civil society to develop more rigorous standards and regulatory measures, grounded in international human rights frameworks, capable of holding Big Tech and other powerful actors to account.

Policy Implications

- Emerging consensus on 'ethical AI' is problematic for its lack of grounding in international human rights law and weak emphasis on accountability and participation. These need to be strengthened so that they can be used to defend the public interest and hold powerful private and public bodies involved in design, development and deployment of AI accountable.
- AI guidelines need to emphasize potential for widening socio-economic inequality, not just discrimination. Capacity and resource constraints in the use of AI enabled technologies is a neglected issue in AI guidelines and debates. These constraints are likely to widen inequalities within and between countries.
- Ethics guidelines that claim to commit to respect human rights should be scrutinized for how well they include the essential principles and standards – anchoring in international human rights legal instruments accountability, participation, privacy, equality. Those that do not do so are 'ethics branding' themselves as committed to human rights.
- Governance of AI-design, development and deployment requires a robust human rights framework, not one that is based on 'ethics' that is an open-ended concept, in order to protect public interest from threats of harmful applications.

With the accelerating pervasiveness of artificial intelligence (AI) in everyday life, the enthusiasm for the opportunities that the technology opens up has been accompanied by a growing public concern about the risks that they pose to individuals and society. When the applications are poorly designed, developed or misused, they can be highly disruptive to both individuals and society. Critics have warned that these new technologies can lead to widening social inequalities, threaten individual human rights, facilitate state authoritarian practices, appropriate and commercialize private data, or create systems that take human autonomy out of decision making. Controversies abound in many fields and country settings. For example, facial recognition technology used in policing can reflect racial bias of the

developer and then recreate discrimination against minorities. It can be used in surveillance and repression of political opposition. Individual personal data – such as patient data in hospital records – can become appropriated and marketed for commercial, political and other purposes. These technologies have deep reach and can transform political, economic and social institutions of the 21st century. Used by and serving the interests of the powerful, whether it is the state or a corporate actor, artificial intelligence's design, development and deployment (AI-DDD) reinforces power structures and can enable oppression of the vulnerable rather than their protection and empowerment. In the context of a neoliberal political economy, Philip Alston, the former United Nations Special Representative on Extreme Poverty and Human Rights, warns of humanity marching 'zombie like' to a dystopian 'digital welfare state' as welfare systems are increasingly automated by digital technologies and used by neoliberal states to surveil, target and punish rather than protect and empower people (Alston, 2019). Technology analyst Zuboff highlights the emergence of 'surveillance capitalism', a new structure of the 21st century economic order built on assets created by surveilling the internet for personal information, and driven by the interests of Big Tech (Zuboff, 2019). In the global context, new technologies have the potential to overcome resource constraints to accelerate poverty reduction, but may well recreate inequalities. Developed in the North for their conditions, and marketed in the South, new technologies for the South may be in undersupply, or inaccessible.

By now, stakeholders in private sector, government and civil society are increasingly engaged in debates about legal and ethical challenges of AI-DDD, in search for value-based norms of 'responsible AI' or 'ethical AI'. While government regulation has been slow to emerge, the risks of untrammeled deployment of AI have catalyzed leading corporations, government agencies, multilateral organizations, professional associations, and human rights groups to issue guidelines aiming to set out ethical principles for responsible AI-DDD. While these guidelines do not carry legal legitimacy or weight, and are issued by individual stakeholders, they are creating narratives and building what amounts to de-facto consensus norms of practice in the field.

The purpose of this paper is to address one of the elements of this debate: the human rights critique of ethics framing in the emerging consensus norms. Critics argue that guidelines framed as 'ethics' render them meaningless; as stated by Alston 'as long as you are focused on ethics, it's mine against yours. I will define fairness, what is transparency, what is accountability. There are no universal standards'. (Alston, 2019). Human rights offer a more robust framework because of their legitimacy as an internationally agreed set of norms, and because they are subject to enforcement – however, soft – through national and international mechanisms (Asaro, 2019; Berthet, 2019; Elsayed-Ali, 2018; HRBDT, n.d.).

This paper explores in detail how human rights are reflected in emergent norms. We review the key principles and standards that are relevant in AI-DDD and scrutinize how they are reflected in stakeholder guidelines. We examine in detail 15 guidelines issued by private, public, professional, and multlistakeholder organizations for their treatment of human rights principles and standards.

1. Stakeholder guidelines for AI-DDD and emerging norms

In just a few years, stakeholder guidelines have proliferated and by April 2020, Algorithm Watch's AI Ethics Global Inventory included more than 160 guidelines, a doubling in the 12 months since the Inventory was first launched (Algorithm Watch, 2019). Several efforts have been made to assess the overall content of these guidelines as a way of identifying key elements of emergent norms. They include Algorithm Watch's ongoing inventory, a review by Jobin and others covering 84 guidelines (Jobin et al., 2019), Harvard University's Berkman Klein Center for Internet and Society's mapping and review of 32 guidelines (Fjeld et al., 2020) and Asaro's (2019) analysis of 28 guidelines from private, civil society and academic actors. These reviews find a set of six to 11 common themes that are present in most guidelines. Moreover, these overlap, though using different terminology, as shown in Table 1. The authors conclude that these trends suggest the convergence towards a set of ethical principles (Jobin et al., 2019), or the emergence of 'a "normative core" of a principle-based approach to AI ethics and governance' (Fjeld et al., 2020, p. 5).

References to human rights is apparently a common feature of existing guidelines issued from across sectors; Berkman Klein found 23 of the 36 documents reviewed referred to human rights. Surprisingly, they also found several government documents (7 out of 13 or 54%) did not refer to human rights (Fjeld et al., 2020). Jobin did not find human rights to be a common theme. This begs the question as to what is meant by the term 'human rights', a question that is explored in the rest of this paper.

Implementation is identified in all these reviews as the critical challenge. The reviews note that the principles are expressed in somewhat vague terms and lack enforcement mechanisms. Jobin and others also note that while there is convergence on principles, guidelines vary significantly on: 'how the principles are interpreted; why they are deemed important; what issue, domain or actors they pertain to; and how they should be implemented' (Jobin et al., 2019, p. 396). Asaro (2019) notes that the guidelines also recognize the need for stronger regulatory mechanisms, and identify examples of guidelines that provide specific and detailed ways in which the principles can be developed into regulatory and policy provisions, pointing out three promising examples.

2. Human rights approach to AI

2.1. AI-DDD as a human rights issue

AI-DDD has emerged as an important area of human rights concern in the last decade. The UN High Commissioner for Human Rights, and UN special rapporteurs have called

Table 1. Common themes identified in reviews of guidelines

Algorithm Watch	Jobin and others	Berkman Klein	Asaro 2020
Human rights	Privacy, freedom and autonomy, trust, dignity	Human Rights; privacy	Respect Human Rights including dignity and privacy
Beneficial to society	Beneficence, sustainability, solidarity	Promotion of human values	Promote human well-being
Accountability	Responsibility and accountability	Professional responsibility; human control of technology; accountability	Ensure responsibility and accountability remain with human designers/operators
Equality and non-discrimination	Justice, fairness and equity	Fairness and non-discrimination	Avoid bias and deception
Transparency	Transparency	Transparency and explainability	Be transparent, reliable and trustworthy
Safety	Non-maleficence,	Safety and security	Do no harm

attention to human rights issues with respect to privacy and freedom of expression (UN General Assembly, 2018; UN High Commissioner for Human Rights, 2018), surveillance and the right to assembly and peaceful protests, (UN High Commissioner for Human Rights, 2020), racial discrimination (Achiume, 2020) and poverty and digital welfare state (which analyses how the increasing automation of the state's distribution of welfare benefits risks denying the poor their rights) (Alston, 2019). The right to privacy, as set forth in the International Covenant on Civil and Political Rights (under Article 17) has been one of the few specific human rights for whose protection there is widespread concern in the digital age (UN General Assembly, 2018). However, while privacy is a major issue, human rights issues are much broader, and concern a wide range of issues related to discrimination, lack of transparency, accountability of business and the state (Achiume, 2020; Alston, 2019; Crawford et al., 2019; OHCHR, n.d.). The human rights community advocates a human rights-based approach to AI-DDD that draws holistically on international human rights.

While human rights advocates often criticize the use of ethical frameworks as lacking universality, guidelines using an ethical framework point to the importance of an ethical AI aligned with society's values 'based on well-founded standards of right and wrong' of which global companies must take account (IBM, 2019). But these two positions are not contradictory, and arguments for cultural relativism have long been addressed by human rights scholars (Donnely, 1984). While human rights principles and standards are universal, their application in different social and cultural contexts varies (Merry, 2006).

A variant of this argument is that human rights law was developed in the predigital age, and may not always have caught up with the challenges posed by rapidly evolving technologies (UN Secretary General's High-level Panel on Digital Cooperation (HLPDC), 2019; High-level Expert Group on Artificial Intelligence (AI HLEG), 2019). As such it carries the limitation of an inability to foresee or anticipate all potential AI harms to society (Latonero, 2018). UNESCO (United Nations Educational, Scientific and Cultural Organization) recently issued recommendation on the ethics of AI which captures this complementarity well:

> ethical values and principles are not necessarily legal norms in and of themselves, [but they] ... can powerfully shape the development and implementation of policy measures and legal norms, by providing guidance where the ambit of norms is unclear or where such norms are not yet in place due to the fast pace of technological development combined with the relatively slower pace of policy responses" (UNESCO Ad Hoc Expert Group (AHEG), 2020)

This implies that human rights norms as currently codified in international, regional and national legal frameworks are not sufficient, but does not imply that these norms are unnecessary.

Moreover, the human rights framework – like all legal frameworks – is not static; it evolves over time, with the constant development of new treaties, and multiple other human rights documentation such as General Comments, and reports of Special Rapporteurs and more. Thus, the UN human rights machinery has started identifying technology in the digital age as an issue and called for studies and debates to develop norms, principles and standards. Work is emerging, particularly from special rapporteurs in area of privacy, freedom of speech, poverty and racial discrimination. (Achiume, 2020; Alston, 2019). Some are already calling for the consideration of new rights; The Rathenau Report, commissioned by the Council of Europe, proposes consideration of two new rights, to reflect the new reality of AI embedded in human life: the right to not be measured, analyzed or coached, and the right to meaningful human contact (Rathenau Institute, 2017).

2.2. Core principles

At the core of the human rights approach is the requirement to apply international human rights normative framework and its laws, principles and standards to AI. The

approach goes beyond regulation of AI to prevent harm. It demands the application of human rights principles in all stages of AI-DDD. Furthermore, the human rights approach addresses the social contexts – particularly the power structures and societal norms – that leave individuals vulnerable to violations. Human rights enforcement is a tool to protect the vulnerable and hold the powerful accountable.

While the core human rights treaties and other instruments establish norms and standards in numerous civil, political, economic social and cultural rights, they also spell out central principles that apply across all these areas; adherence to these principles is particularly relevant to the way that human rights are implemented (UN, 1966). These principles include:

- universality – human rights are universal and apply to all people in the world. They reflect a universal consensus on ethical values and principles regardless of cultural context;
- right to equality and non-discrimination – human rights apply to all individuals by virtue of being a human. All individuals have equal rights and the right to equality and non-discrimination is a core principle;
- participation – people have a right to participate in civic life, particularly in decisions that affect their life. This requires rights bearers access to information and transparency on the part of duty bearer; and
- accountability and remedy - human rights norms are not only about the enjoyment of rights by individuals but also the accountability of correlate duty bearers.

The human rights regime is also clear in defining the nature of obligations of duty bearers. Not only are rights what humans are entitled to, these entitlements impose correlate obligations on the state and other powerful actors who can ensure that people's rights are realized. These obligations are not only to *respect*, but to take proactive steps to *protect* rights from being violated by others and to *fulfill* these rights by taking proactive efforts (CESCR, 1990). While the states are the primary duty bearers, businesses also have obligations, as spelt out in the Guiding Principles for Businesses and Human Rights (OHCHR, 2011).

AI-DDD as currently practiced raise significant concerns with respect to these principles, and the right to privacy. We discuss each of these in turn.

2.3. Equality and non-discrimination

Bias is a major issue in AI-DDD. Algorithms are inherently biased. Design and development reflect the judgement of humans involved while data collection, interpretation and dissemination can never be neutral, often excluding or misrepresenting, racial and gender minorities (Crawford et al., 2019) (Panch, 2019). These issues have become well known with respect to facial recognition used in policing that discriminates sharply against racial minorities in the US (Achiume, 2020; Myers et al., 2019). They are also an issue in health, for example, in estimating health risks that determine allocation of resources for care to patients, or the predictive accuracy of prescribed treatment (Ferryman and Pitcan, 2018; Obermeyer et al., 2019). The potential and actual discriminatory impact on people affected by AI decision-making in access to health and other social entitlements is well documented (Alston, 2019; Eubanks, 2018).

Algorithmic bias also reinforces existing inequalities, because bias often reflects existing social norms that discriminate against minorities and women. AI compounds this discrimination through unequal access, which further excludes the marginalized from the benefits of AI technology (Achiume, 2020; Smith et al., 2020). Investments in new technologies are financed commercially, even when they are used for public purposes such as healthcare, education, social protection, and security, and can be misaligned with public priorities, or the needs of low and middle income countries (LMICs). The digital divide may consistently exclude children, minority populations and populations in developing countries; these populations may not even be producing data, or what they produce may not be properly captured as training data needed to ensure the predictive accuracy of the algorithm, or be irrelevant to the problem the application is purported to solve (Erikson, 2018).

Infrastructure constraints continue to be significant in LMICs; in some countries, between 80% and 90% the population lack access to the internet (Internet World Statistics, 2020). While mobile phones can compensate for some of that lack, unequal phone ownership and cell-coverage introduce inequalities of their own. Further, AI applications developed in high income countries tend not to consider social factors necessary for the application to function properly in LMIC contexts. While there is great potential for AI to reduce inequality within and among countries, there is an ever-present risk that it will instead aggravate already high rates of inequality. The dynamics of AI and inequalities lies not only in the technology but in the unequal social and economic structures in which AI-DDD takes place (Ferryman and Pitcan, 2018; Smith et al., 2020).

2.4. Participation

Participation is a core human rights principle; all people are entitled to meaningful, informed participation in the decisions that affect them. This includes participation in AI-DDD with potential impact on the lives of the application's users or subjects. Participation ensures that applications are responsive to the needs of the people they are meant to benefit and that they do not produce harm (intentional or unintentional) to that population. Participation is also the means by which rights-holders can hold public and private actors accountable for the impact of AI on their well-being.

Participation is meaningless without access to information or freedom of expression. The lack of transparency in AI-DDD, which mostly takes place in the private sector, is a major human rights issue. Among the obstacles are companies' protection of proprietary information (such that meaningful participation is not possible), an absence of platforms for the review of AI pilot applications, the asymmetry of

knowledge and power between developers and users regarding how algorithms work, what data protections are necessary, the sources of training data, the evidence of impact, etc. Collectively, these obstacles impede participation and accountability, and thus society's trust in AI.

2.5. Accountability and remedy

Accountability is the means through which rights are actually realized and consists of three interdependent elements: *responsibility* – who/what institution has the duty to perform the respect, protection and fulfillment of rights, and at what standard; *answerability* – a formal process of transparency whereby the public can demand and receive answers to questions about how those in authority reached their decisions; and finally, *enforceability* – when human rights standards are violated and individual or community harm results, a mechanism exists to sanction those responsible *and* provide a remedy to the persons affected (OHCHR, 2013). Key to the Human Rights principle of accountability is the critical element of *remedy* to put wrongs right.

The demand for accountability and remedy is the cornerstone of human rights practice and sets this approach apart from other efforts to promote ethical AI-DDD. It addresses the power imbalance that is pervasive in human rights violations; human rights protection is most often about defending the powerless against egregious behavior of the powerful. The risks to human dignity from AI result most often from the context in which it is being designed, for whom and by whom it is being developed, and for what purpose it is deployed. While public research in universities and government institutions played a dominant role in advancing science and technologies for health, the new technological advances are largely led by private finance, corporations and start-ups, and philanthropies. As the Human Rights, Big Data and Technology project explains, 'the presence of tech giants in this sector and the increase of public-private data sharing could potentially enhance inequalities in health care provision. This is contrary to human rights "right to health" legislation' (HRBDT, n.d.).

2.6. Privacy

Privacy, data protection and ownership are major human rights concerns in the AI age. The right to privacy is defined in Article 17.1 of the International Covenant on Civil and Political Rights as 'No-one shall be subjected to arbitrary or unlawful interference with his privacy, home or correspondence'. The European Convention for the Protection of Human Rights and Fundamental Freedoms define the right in Article 8.1 'Everyone has the right to respect for his private and family life, his home and his correspondence'.

The Rathenau Institute summarizes the core privacy problem thus 'The primary business model of the Internet is built on mass surveillance', and it quotes Professor Julie E. Cohen 'We the citizens have been reduced to raw material-sourced bartered and mined in ... "privatized commons" of data and surveillance' (Rathenau Institute, 2017). Or as the statement of a multi-national radiology community explains, 'because developing AI-driven machines today requires massive amounts of well-labeled radiology data, the value of those data is skyrocketing and the drive to provide commercial access to radiology data will become overwhelming' (American College of Radiology et al, 2019). Actual and potential violations to the right of privacy are the natural consequence of mass surveillance, and the difficulty people have knowing, let alone controlling, what personal data are being shared for what purpose. Personal data collected in the health system have become open to use, dissemination, and monetization without consent (for example, Ebeling, 2016; Gopichandran et al., 2020). Violation of the right to privacy can have consequences for the realization of rights to health, to education, to equal justice or protection from violence and numerous other rights. But the more far-reaching concern is that systems for privacy and consent over data are no longer sufficient. In the age of internet surveillance, a few tech giants monopolize access to personal data and monetize them for a wide range of purposes. For this reason, privacy is given particular prominence in debates on ethical AI-DDD and is included in this paper's review, though it is not one of the core human rights principles.

These concerns in the practice of human rights in AI are inter-related and shaped by the context of the social, legal, financial and political structures within which AI-DDD takes place. Bias is driven by the existing structures of socio-economic inequalities; participation is hampered by the power of the tech industry to resist transparency; companies are not accountable to rights-holders because of the weakness of national and international governance; and privacy is not protected because of the power of financial interests in monetizing private information, and the inability of individuals to assert their data ownership, and hence their right to privacy. This creates misalignment between public priorities, including for global health, and investment priorities driven by profit motives.

2.7. Obligations of duty bearers

Human rights entail correlate obligations on the part of the state and other powerful actors. These obligations are not only to governments; businesses and trans-national actors have obligations not only to respect rights but to proactively protect and fulfill the provisions laid out in the international human rights treaties, and to practice the principles of non-discrimination, participation and accountability in all their actions. By fulfilling their obligations, duty-bearers undergird the human rights framework and ensure its application to AI-DDD, and make available accountability mechanisms which allow the powerless to ensure duty-bearers meet their obligations and protect themselves from violations.

3. Review of stakeholder guidelines

As noted above, many of the stakeholder guidelines refer to 'human rights' yet critics consider that they are a 'human

right free zone' (Alston, 2019). How do the guidelines (listed in the Appendix) that were considered to have made an attempt to address the human rights concerns of AI-DDD actually do so? The 15 included those issued by government, private sector, professional associations, civil society organizations. Some of the guidelines covered AI as one digital technology among several (UN Secretary General's High-level Panel on Digital Cooperation (HLPDC), 2019), while others limited guidance to a specific subset of AI (Amnesty International and Access Now, 2018), but all aimed to contribute to the ethical development of artificial intelligence.

Our analysis looked for evidence that the guidelines adopted, in whole or in part, a human rights framework as defined above: first and foremost, anchored in international or regional human rights law, and second, incorporating proposals for AI-DDD that operationalize key human rights principles: equality and non-discrimination; participation of rights holders; accountability that includes responsibility, answerability/transparency and enforceability/remedy. The obligation of duty-bearers to promote, protect and fulfill rights is understood to be embedded in these principles, and therefore, for the purposes of this paper, will not be the subject of a separate analysis. Because of particular human rights concerns over privacy, evidence for the protection of privacy is an additional element of the analysis.

3.1. Anchored in international and regional international human rights law

Of the 15 guidelines reviewed, seven explicitly referenced human rights instruments and procedures as the best framework for assessing AI-DDD's potential benefits and harms. These include: European Commission for Europe (European Commission, 2018); European High Level Expert Group on AI, Ethics Guidelines for Trustworthy AI (High-level Expert Group on Artificial Intelligence (AI HLEG), 2019); Amnesty International and Access Now, The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems(Amnesty International and Access Now, 2018); IEEE, Ethically Aligned Design (IEEE, 2019); UN Secretary General's High Level Panel on Digital Cooperation (HLPDC), The Age of Digital Interdependence (HLPDC, 2019); Access Now, Human Rights in the Age of AI (Access Now, 2018) and Microsoft, Global Human Rights Statement (Microsoft, 2019).

The European Commission affirms its approach to AI ethics as based on fundamental rights, and that EU and international human rights treaties 'provide the most promising foundation for identifying abstract ethical principles and values' (High-level Expert Group on Artificial Intelligence (AI HLEG), 2019). The Toronto Declaration focuses on operationalizing the principle of non-discrimination, and states 'As discourse around ethics and artificial intelligence continues, this Declaration aims to draw attention to the relevant and well-established framework of international human rights law and standards' (Amnesty International and Access Now, 2018). IEEE establishes the respect for human rights as its very first general principle: 'A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights', and human rights considerations are reflected across EAD's nine chapters (IEEE, 2019). However, the title of its publication *Ethically Aligned Design,* points to an embrace of both frameworks in pursuit of A/IS that benefits humanity; it joins other guidelines in recognizing that approaches based on ethics and on human rights reinforce each other.

The other guidelines use ethics as the framework for determining AI's benefit to society, but include respect for human rights as one ethical principle among several; its primacy is not established. Where human rights are referenced, they tend to be nested within an ethics framework; the Montreal Declaration lays out principles that 'offer an ethical framework that promotes internationally recognized human rights in fields affected by the rollout of AI' (Montreal Declaration Responsible AI, 2018). Smart Dubai (2019) firmly grounds its guidelines in ethics but includes a co-equal principle that 'AI should be beneficial to humans and aligned with human values'. IBM and Google refer to human rights even more opaquely: 'Ethics is based on well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues.' (IBM, 2019); Google declares 'We will strive to make . . . information readily available using AI, while continuing to respect cultural, social, and legal norms in the countries where we operate' (Google, 2019). Surprisingly, WHO's draft strategy makes only a fleeting reference to human rights, mentioning violations as one of the pitfalls of AI use (WHO, 2019).

Microsoft is unique in that its *Global Human Rights Statement* is very much grounded in international human rights convention and commits the company to respecting and promoting human rights. Yet at the same time, Microsoft's principles for '*Responsible AI*' commits it to 'advancement of AI driven by ethical principles that put people first'. Respect and promotion of human rights is not evident in *Responsible AI's* six principles of fairness, reliability and safety, privacy and security, inclusiveness, transparency and accountability. It is not clear if or how these guidelines operate simultaneously at Microsoft.

Four arguments are given for the use of the human rights framework: first that the international system of human rights law is established, operational, virtually universal, thus 'particularly well suited to borderless technologies' (Amnesty International and Access Now, 2018); second that national, regional and international human rights mechanisms are also uniquely equipped to deal with restitution and remedy (Amnesty International and Access Now, 2018; AI HLEG, 2019); third that the human rights framework commands legitimacy as an internationally agreed set of norms that 'represent the most universal expression of our shared values' (Berthet, 2019), and fourth, that it is equipped to deal with power-imbalances (Access Now, 2018).

3.2. Equality and non-discrimination

All of the documents reviewed addressed the principle of equality and non-discrimination, although not always as a

human rights consideration. To prevent unintentional discrimination and bias, the guidelines variously called for the avoidance of 'unfair bias' (Google, 2019) and of non-representative training data (Smart Dubai, 2019), and 'freedom from bias' (High-level Expert Group on Artificial Intelligence (AI HLEG), 2019). Some guidelines went further by calling for disclosure of inherent bias and discrimination through impact assessments (Smart Dubai, 2019), investigations of bias in the data and audits to capture unintentional bias (IBM, 2019). One guideline called for states to update existing measures to prevent discrimination and to address the risks posed by machine learning (Amnesty International and Access Now, 2018). Several guidelines also recognized the importance of preventing 'unjust impacts' (IBM, 2019), of ensuring that no AI application reinforce or reproduce discrimination (Montreal Declaration Responsible AI, 2018) and of establishing a 'right not to be subject to AI-decision making' (Madiega, 2019). There was also recognition that 'harms related to the use of AI disproportionately affect marginalized populations due to input data that is not representative' with the accompanying risk of baking in inequalities (Access Now, 2018). IEEE not only called for AI-DDD to be undergirded with 'principles of nondiscrimination, equality, and inclusiveness', but for particular attention to 'be given to vulnerable groups, to be determined locally, such as minorities, indigenous peoples, or persons with disabilities' (IEEE, 2019).

As much as combatting discrimination, the guidelines reviewed called for AI to advance equality, with access guaranteed to all, ensuring that 'everyone benefits' (Montreal Declaration Responsible AI, 2018), equal respect of human dignity (High-level Expert Group on Artificial Intelligence (AI HLEG), 2019) and 'leaving no-one behind' (HLPDC, 2019; (Microsoft, 2019). Several guidelines called specifically for inclusion and diversity in AI-DDD in order to off-set unintentional bias, sensitivity to a wide range of cultural norms (IBM, 2019), and use of local languages (Microsoft, 2019). The radiologists in their European and North American Multisociety statement, and health professionals in the American Medical Association (AMA) emphasize the need for proactive steps to address bias. AMA commits to promote health care AI that ... 'identifies and takes steps to address bias and avoids introducing or exacerbating health care disparities, including when testing or deploying new AI tools on vulnerable populations' (AMA, 2018a, 2018b). Radiologists and WHO address resource inequalities facing resource poor hospitals (American College of Radiology et al 2019), and resource poor countries (WHO, 2019). Radiologists recommend users of AI tools be aware of the potential bias against resource-poor communities due to their underrepresentation in the data used in training and testing models, ensuring diverse composition in committees monitoring AI tools, and setting equal access to AI tools as a goal (American College of Radiology et al, 2019). WHO recognizes, as one of the four guiding principles, the major impediments faced by least developed countries (WHO, 2019).

Civil society and multi-stakeholder sources emphasize that a human rights framework focuses attention on addressing power imbalances in AI-DDD (Access Now, 2018; Latonero, 2018). However, only the Toronto Declaration called specifically for establishing procedures which would provide a remedy to anyone for whom an AI application had a discriminatory impact: 'harms from discrimination have to be immediately addressed throughout the [AI] system's lifecycle' and states should 'document actions taken to mitigate discrimination' (Amnesty International and Access Now, 2018). In the absence of remedy, the harm continues to affect the individual or group by, for example, denying them state benefits to which they are entitled (Alston, 2019), denying them a just criminal sentence or simply by preventing their equal access to goods and services available to others.

Thus, recognition of the human right principle of equality and non-discrimination, in particular the risk of bias in both data and AI design, was quite robust in the guidelines reviewed. They varied in the degree to which specific guidance was given on proactive measures to be taken to address bias, social inequalities and discriminatory impact.

3.3. Participation

Most of the guidelines under review made mention of the need to consult stakeholders at some point in the AI-DDD process (although few recognized participation as a human right), and a number made proposals for rectifying the power imbalance between AI designers and users/subjects. For a subset of guidelines, stakeholder consultation was framed in the context of market research: 'AI subjects and society should be consulted to inform AI development' (Smart Dubai, 2019); 'collect feedback from users to correct bias' (IBM, 2019); and 'engage stakeholders and rights holders to obtain input to help evolve our approach' (Microsoft, 2019).

However, the majority of documents recognized that stakeholder engagement, at whatever stage of the AI-DDD life-cycle, required special efforts to address the information asymmetries impeding meaningful participation. This emphasized both educating the public as well as transparency by those engaged in AI-DDD. Among the calls for meaningful participation are: 'Educate the public on the benefits and challenges of AI' and 'empower rights-holders to claim and exercise rights' (IEEE, 2019); 'it is crucial to empower citizens to develop critical thinking on AI' (Montreal Declaration Responsible AI, 2018); 'break down barriers to information and bring diverse voices to the table' through the creation of mechanisms to ensure equitable participation of women, marginalized and LMIC populations (HLPDC, 2019); and 'enable end-users and the broader society to be informed' such that they can insist upon respect for the principles of trust-worthy AI (AI HLEG, 2019).

The connection between participation and accountability was identified in some guidelines, with a call for the establishment of mechanisms to provide external feed-back on rights violations (Madiega, 2019) or to which errors could be reported (Montreal Declaration Responsible AI, 2018) and for engagement with local stakeholders to mitigate any potential human rights impacts (Microsoft, 2019). There was also a

recognition that the complexity of AI means that ordinary citizens may need to participate through the establishment of trusted independent bodies which certify to the public that an AI system is transparent, accountable and fair (AI HLEG, 2019), or through civil society organizations with the capacity to conduct human rights due diligence for and of companies (Microsoft, 2019).

While several guidelines called for national policies and business regulations of AI to be founded on a rights-based approach, only four explicitly identified participation as a key principle. The Montreal Declaration stipulated *democratic participation* as one of its ten principles and focused particularly on the need for transparency. The Declaration calls for 'citizens to have the opportunity and skills to deliberate the social parameters and limits' of any public AI deployment which would significantly impact citizens' lives (Montreal Declaration Responsible AI, 2018). IEEE states that AI developers should 'Encourage and support a high degree of participation of duty bearers, rights holders, and other interested parties' and 'empower rights holders to exercise and claim their rights' (IEEE, 2019). Ethically Aligned Design also recognizes how power imbalances affect the equitable participation of stakeholders and recommends a methodology for mitigating this imbalance. The Toronto Declaration and Access Now made mention of the right to participation in passing.

Thus, the guidelines reviewed recognize the importance of stakeholder engagement, which can be considered a proxy for the human right to participation, while only two, The Montreal Declaration and Ethically Aligned Design, elaborated on participation as a right and a guiding principle. Some also recommend the establishment of mechanisms that correct some of the information and power asymmetries which impede meaningful participation in AI-DDD.

3.4. Accountability and remedy

All of the guidelines reviewed included accountability among the ethical and human rights principles with which they proposed to govern AI-DDD. However, most emphasized transparency; the extent to which accountability includes the key elements of *responsibility, answerability and enforceability*, as defined by OHCHR, varied greatly. Examination of how accountability is defined in guidelines suggest each views the term's meaning, and hence its operability, quite differently. For the majority of guidelines, accountability for the impact of AI-DDD is limited to establishment *of responsibility* for the design and development process, so that how and why an AI system acts is explainable and auditable. Hence guidelines call for participants in the AI-DDD process to maintain a decision journal and apportion clear responsibility among these participants (Smart Dubai, 2019); to document and continually evaluate trade-offs (AI HLEG, 2019), recording the processes, people involved and their roles/responsibilities, so as to ensure that decisions are discoverable and traceable, embedding the equivalent of a flight data recorder in each AI system (IEEE, 2019); ensuring everyone involved in the AI-DDD process understands their own responsibility *and* where it ends (IBM, 2019). Human oversight and control are emphasized as critical to accountability across the guidelines reviewed. The clarification of roles and responsibilities of participants in AI-DDD and the documentation of decision processes create the basis for answerability-and transparency-thought external audits.

Generally, however, the accountability mechanisms by which users/subjects could take these audits to demand an explanation (let alone remedy) are not elaborated. Guidelines call for transparency through external audits, with public release of results in order to increase trust in AI systems (Smart Dubai, 2019), for independent audits with the protection of company whistleblowers (AI HLEG, 2019), oversight by third parties and certification requirements (HLPDC, 2019), public sharing of errors and flaws on a global level (Montreal Declaration Responsible AI, 2018) and public disclosure when a machine-learning AI application is being used in the public sphere (Amnesty International and Access Now, 2018). Thus, for the majority of guidelines reviewed, for AI systems to be accountable, the people responsible for their design, development and deployment must be identifiable, the decisions they make documented, with the final product's operation in the world subject to external audit whose results are shared with the public. Even IEEE's *Ethically Aligned Design* which includes accountability as one of its general principles (and states that 'A/IS regulation, development, and deployment should. . .be based on international human rights standards and standards of international humanitarian laws') limits accountability's scope to responsibility and answerability: 'Oblige states, as duty bearers, to behave responsibly, to seek to represent the greater public interest, and to be open to public scrutiny of their A/IS policies'. At one point, *Ethically Aligned Design* does, however, call for manufacturers of these systems to be accountable 'in order to address legal issues of culpability and . . . apportion culpability among responsible creators (designers and manufacturers)'which suggests an interpretation of responsibility which may lead to remedy. (IEEE, 2019).

The responsibility component of accountability is overwhelmingly mentioned in the guidelines, and that of answerability somewhat so, in that publicly available audits also imply a company's willingness to both be transparent and answer for its actions.

With the exception of IEEE's *Ethically Aligned Design* (which can be considered to have partially addressed accountability with remedy), the remaining six guidelines which adopted human rights law as the framework for ethical AI also embedded enforceability, restitution or remedy within the accountability principle for harms potentially caused by an AI application. In addition, Smart Dubai while using an ethics framework, called for an appeals procedure to challenge a decision made by an AI system, as well as 'consideration' of compensation for AI subjects 'inconvenienced' by AI decisions. (Smart Dubai, 2019); this points to a recognition that remedy may be needed even though 'consideration' is far from a right to remedy. The European Commission's High-level Expert Group insisted upon AI with *accessible* redress mechanism and remedies, so that

users/subjects would have trust that when things go wrong, accountability and redress would follow; the HLEG also called for particular attention to be paid for vulnerable persons or groups (AI HLEG, 2019). Microsoft, alone among the major private sector companies, includes in its Global Human Rights Statement the enforceability component within its commitment to the principle of accountability: 'Ensuring accountability by providing effective grievance mechanisms and access to remedy in situations where Microsoft may have caused or contributed to an adverse human rights impact' (Microsoft, 2019). The Toronto Declaration and Access Now call for using the human rights mechanisms to both monitor human rights impacts and ensure redress and remedy (Amnesty International and Access Now, 2018). The UN Secretary General's Digital Cooperation panel further recognized that human rights treaties themselves, adopted in the pre-digital age, may need to be updated, and invited 'views from all stakeholders on how human rights can be meaningfully applied to ensure that no gaps in protection are caused by new and emerging digital technologies' (HLPDC, 2019).

3.5. Privacy

All guidelines, without exception, address privacy and data protection concerns. However, only a few address these concerns as a human rights matter; of those that do, all but one (Smart Dubai) adopted the human rights framework. The UN Panel's Age of Digital Cooperation recognizes the right to privacy, and the urgency of finding models which 'share the value extracted from personal data with the individuals who provide it' (HLPDC, 2019). Privacy and data protection is one of European Commission's seven principles for trustworthy AI, which principle 'includes respect for privacy, quality and integrity of data and access to data' and further calls for a guarantee to privacy and data protection throughout the system's entire life cycles (AI HLEG, 2019). IEEE recommends establishment of governance frameworks for AI-DDD which 'ensure A/IS does not infringe on human rights, freedoms dignity and privacy' and that AI-DDD creators 'shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity (IEEE, 2019). Access Now made the clearest statement 'Privacy is a fundamental right essential to human dignity, it reinforces other rights such as freedom of expression and of association' (Access Now, 2018). Yet Access Now also expressed enormous concern that with 'people creating a trail of data for every aspect of their lives', there is a risk to privacy when pieces of data are aggregated so extensive that 'it is questionable whether data protection is even possible' (Access Now, 2018). However, the guideline goes on, data protection laws could 'go a long way to addressing the many human rights risks posed by AI' (Access Now, 2018).

Smart Dubai was the only ethics guideline that referred to privacy as a right; under its principle 'We will respect people's privacy' is stated that privacy should be respected and surveillance systems should not be deployed is they violate UAE's 'accepted standards of privacy, human dignity and people's rights'. (Smart Dubai, 2019) The Montreal Declaration establishes a privacy and intimacy principle such that people must be protected from AI intrusion and data acquisition and archiving systems. (Montreal Declaration Responsible AI, 2018) Google calls for the incorporation of privacy design principles (Google, 2019) and IBM for the protection of user data and their power over its access and use such that 'privacy settings [are] clear, findable and adjustable' (IBM, 2019).

3.6. Overview of findings

The main findings are summarized in Table 2 and discussed below. The 15 guidelines reviewed in this paper are but a tiny subset in the universe of the rapidly proliferating stakeholder guidelines for AI, and include 12 identified as reflecting human rights. Yet they reveal some key issues about the incorporating human rights principles.

The recognition of the potential human rights violations, or harms of AI, and the importance of preventing harm, is accepted and recognized in all 15 guidelines. However, using the human rights framework – as international law – as a means for preventing such harm is adopted by only seven of the 15 guidelines examined, and all seven of these incorporate principles of non-discrimination, participation, accountability and remedy in a holistic manner. (We include *Ethically Aligned Design* here, since its embrace of accountability with remedy is at least partial). They also provide guidance on operationalizing their recommendations (of the 7, only Microsoft's Global Human Rights Statement fails to do so). For example, less than one year after issuing its report, the UN High-level Panel on Digital Cooperation released, through the Secretary General, a roadmap for implementation of its recommendations (UN Secretary General, 2019). The EC's High-level Expert Group goes as far as to provide checklists to help public and private actors take action to advance trustworthy AI (AI HLEG, 2019). However, even these efforts do not go far enough to provide enforceable mechanisms for remedy.

The remaining guidelines adopt an ethics framework, approaching human rights in a perfunctory or piecemeal fashion. While equality, non-discrimination and privacy are included in most guidelines (though not always as human rights), the principles of participation and accountability are either addressed inadequately or misinterpreted. Operational direction is also weak to non-existent in the guidelines using an ethics framework. As compared to guidelines issued by all other stakeholders reviewed, those produced by the private sector were least likely to have incorporated elements of a human rights framework (although Microsoft stands out as a company that attempts to do so).

Unlike the human rights principles, privacy is universally recognized as essential in the set of guidelines reviewed, whether it is viewed as a human right or a data protection concern or both. However, only six of 15 guidelines

Table 2. Core human rights principle in selected stakeholder guidelines

	Organization	Ethics framework	International Human Rights framework	Equality and Nondiscrimination# Bias	Participation (as HR principle)	Participation ("consultation, education")	Accountability and Remedy (*Responsibility, Answerability, Enforceability)	Privacy (general)	Privacy as an HR Issues
Public	Smart Dubai	Yes		Yes		Yes	R, A	Yes	Yes
	Euro Parliament		Yes	Yes#		Yes	R, A, E	Yes	
	HLEG Commission		Yes	Yes#		Yes	R, A, E	Yes	
	HLP-DC		Yes	Yes#		Yes	R, A, E	Yes	Yes
	WHO Strategy	Yes		Yes			R	Yes	
Corporate	IBM	Yes		Yes		Yes	R, A	Yes	
	Google	Yes		Yes		Yes	R, A	Yes	
	Microsoft		Yes	Yes#		Yes	R, A, E	Yes	
Civil Society	Amnesty International and Access Now		Yes	Yes#	Yes	Yes	R, A, E	Yes	
	Access Now		Yes	Yes#	Yes		R, A,E	Yes	Yes
	UNI Global Union	Yes		Yes			R, A	Yes	
Professional Association	American Medical Association	Yes		Yes			[A]!	Yes	
	American College of Radiology and others	Yes		Yes			R	Yes	Yes
	IEEE	Yes	Yes	Yes#	Yes	Yes	R, A {E}*	Yes	Yes
Multistakeholder	Montreal Declaration	Yes		Yes	Yes	Yes	R,A	Yes	
				# Right to Non-discrimination			* Partial ! liability		

explicitly recognized privacy as a right. Neither ethics nor human rights guidelines on AI-DDD have as yet identified how that privacy and data protection can be fully operationalized.

While our review of 15 guidelines finds, consistent with other studies, that individual bias is addressed, less attention is addressed to the need to use new technologies to redress inequalities, either within or between countries. This could be done by addressing the digital divide, by building capacities, by investing in pro-poor technologies, and by holding services accountable for potentially discriminatory use of AI enabled health technologies that entrench inequalities.

Finally, of all the core principles, accountability is the most problematic. The guidelines using ethical frameworks focus on 'responsibility', 'transparency' and 'liability', ignore remedy, and omit mention of answerability and enforcement.

4. Conclusions

One of the key differences between an ethics approach and a human rights approach is the way accountability and power asymmetries are addressed. Human rights are structured to be not only a set of values but designed to operate as a check on arbitrary power. AI development is dominated by 'Big Tech' and other businesses whose power rival or surpass those of the state, and overwhelming the power of citizens. AI applications are developed as commercial products and deployed by private and public entities. Guidelines are intended to protect citizens from harmful decision making and abusive use of AI by the state or by private businesses that infringe on their freedom and dignity.

Accountability – in its full sense of the concept that goes beyond transparency – is thus central to the very purpose of the guidelines. Participation – including the related requirement of information transparency - and accountability are inter-related principles that build on each other in the practice of human rights; it is only when people have the information and can participate in decisions that the designers and users of AI-DDD can be held to account. Without a robust commitment to accountability including answerability and enforceability, and related conditions of transparency and participation, ethical guidelines are rendered toothless. Indeed, Asaro (2019, p. 7) contends that 'by and large, the ethical principles produced by corporations appear to have as a primary purpose to foster a brand image for the company as socially benevolent and trustworthy', and also to forestall the introduction of legally binding regulation. There is a great deal of skepticism about the guidelines, often alleged to be 'ethics washing', while 'Few companies can show tangible changes to the way AI products and services get evaluated and approved' (Hao, 2019, p.1).

Algorithm Watch, which developed a global inventory of 160 guidelines, in noting their often vague formulation, questioned whether guidelines 'that can neither be applied nor enforced are not more harmful than having no ethical guidelines at all' (Algorithm Watch, 2019). Of the private sector guidelines reviewed here, only Microsoft's Global Statement fully commits to respecting and promoting human rights across its policies, supply chains, partnerships and workforce, and to ensuring 'access to remedy where Microsoft might have contributed to and adverse human rights impact' (Microsoft, 2019). However, as noted above, this statement competes with the company's ethics framework for 'Responsible AI'.

Accountability is also related to principles that have been clearly defined and codified; the weakness of an ethical framework is that its principles are open to interpretation, and lack any institutional mechanism for enforcement.

This review attempted to define a human rights framework for AI-DDD and applied that definition to 15 guidelines, all of which had been selected because they made some attempt to address human right concerns. The findings make clear that application of a more rigorous human rights approach highlights the gaps in enforceability of guidelines adopting an ethics framework, and as such the inability of society and its vulnerable members to hold duty-bearers to account for meeting their human rights obligations. In addition, consistent with their vague formulation, the ethics guidelines reviewed offered few if any mechanisms for implementation, and as such failed to provide any realistic means for shaping the design, development and deployment of AI in the real world.

Adherence to the human-rights framework of AI-DDD can have real life consequences. For example, in health, without eliminating bias in the data set used in developing applications, whole sections of the global society risk being excluded from benefiting from technological advances, furthering health inequalities. If marginalized populations are not included in AI-DDD participation, their crucial input on morbidity and mortality risks or socio-cultural determinants of health will be missing, opening the potential for discrimination in health care. In the absence of accountability mechanisms, AI applications can generate decisions with life and death consequences for vulnerable and marginalized populations. In the absence of robust privacy and data protection, data breaches can affect a patient's access to care or to employment.

While human rights and ethics are not mutually exclusive but are complementary, the real value added of human rights resides in accountability in its full sense, including remedy. The normative content of the guidelines needs to be set in the context of economic incentives and power asymmetry in which AI-DDD takes place: that it is a market driven, commercial process, driven by technology companies that have enormous power, with global reach. Mechanisms for holding these powerful actors accountable is ultimately what the ethical frameworks need to buttress. The framework of human rights and the human rights principles which operationalize those values and norms, command political consensus as universal values, associated with national and international legal machinery. This framework is therefore the one that is 'fit for purpose' (Latonero, 2018). It is therefore essential to bring the language of human rights and emphasis on accountability to these stakeholder guidelines which reflect consensus thinking, and reinforcing them by creating a narrative of what is meant by 'responsible AI'.

References

Access Now. (2018) *Human Rights in the Age of Artificial Intelligence*. s.l.: Access Now. Available from: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf [Accessed 19 July 2020].

Achiume, E. T. (2020) *Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis – Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance*. [Online] Available from: Available from: https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session44/Pages/ListReports.aspx [Accessed 19 July 2020].

Algorithm Watch. (2019) *AI Ethics Guidelines Global Inventory*. [Online] Available from: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/ [Accessed 5 March 2020].

Alston, P. (2019) *Report of the Special Rapporteur on Extreme Poverty A/74/48037*. New York: United Nations.

AMA. (2018a). *Augmented Intellignece in Health Care H-480.940*. [Online] Available from: https://policysearch.ama-assn.org/policyfinder/detail/augmented%20intelligence?uri=%2FAMADoc%2FHOD.xml-H-480.940.xml [Accessed 21 April 2021].

AMA. (2018b). *Policy Recommendations on Augmented Intellligence in Health Care H-480.940*. s.l.: AMA. Available from: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf [Accessed 19 July 2020].

American College of Radiology, European Society of Radiology, Radiology Society of N. America, Society for Imaging Informatics in Medicine, European Society of Medical Imaging Informatics, Canadian Assoc of Radiologists et al. (2019) *Ethics of AI in Radiology: European and North American Multisociety Statement*, s.l.: s.n. Available from: https://www.acr.org/-/media/ACR/Files/Informatics/Ethics-of-AI-in-Radiology-European-and-North-American-Multisociety-Statement--6-13-2019.pdf [Accessed 21 April 2021].

Amnesty International and Access Now. (2018). *Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems*. [Online] Available from: https://www.torontodeclaration.org/declaration-text/english/ [Accessed 4 July 2020].

Amnesty International and Access Now. (2018) *The Toronto Declaration: Protecting the Right to Equality and Non-DIscrimination*. s.l.: s.n. Available from: https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/ [Accessed 21 April 2021].

Asaro, P. (2019) *A Review of Private Sector AI Principles: A Report Prepared for UNIDIR*. Geneva: unpublished report prepared for UN. Institute for Disarmament Research.

Berthet, A. (2019) *Why do Emerging AI guidelines Emphasize "Ethics" over Human Rghts?*. [Online] Available from: https://www.openglobalrights.org/why-do-emerging-ai-guidelines-emphasize-ethics-over-human-rights/ [Accessed 28 June 2020].

CESCR (1990) *General Comment 3: The Nature of State Parties' obligations*. New York: United Nations.

Crawford, K., Whittaker, M., Dobbe, R., Fried, G., Green, B. et al. (2019) *AI Now Report 2019*. [Online] Available from: https://ainowinstitute.org/AI_Now_2019_Report.html. [Accessed March 2020].

Donnely, J. (1984) 'Cultural Relativism and Universal Human Rights', *Human Rights Quarterly*, 6 (4), pp. 400–419.

Dubai, S. (2019) *AI Ethics, Principles and Guidelines*. Dubai: Smart Dubai.

Ebeling, M. (2016) *Healthcare and Big Data: Digital Specters and Phantom Objects*. New York: Palgrave Macmillan.

Elsayed-Ali, S. (2018) *New Human Rights Principles on Artificial Intelligence*. [Online]. Available from: https://www.openglobalrights.org/new-human-rights-principles-on-artificial-intelligence/ [Accessed 28 June 2020].

Erikson, S. (2018) *Cell Phones ≠ Self and Other Problems with Big Data Detection and Containment during Epidemics*, [Online] Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6175342/ [Accessed 5 March 2020]

Eubanks, V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. New York: St. Martins Press.

European Commission. (2018) *AI for Europe*. Brussels: European Commission.

Ferryman, K. and Pitcan, M. (2018) *Fairness in Precision Medicine*. New York: Data & Society.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M.. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. [Online] Available from: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420 [Accessed 6 March 2020].

Google. (2019) *Artificial Intelligence at Google: Our Principles*. Available from: https://ai.google/principles/ [Accessed 21 April 2021].

Gopichandran, V., Ganeshkumar, P., Dash, S. and Ramasamy, A. (2020) 'Ethical Challenges of Digital Health Technologies: Aadhaar, India', *Bulletin of World Health Organization, Issue*, 98 (4), pp. 277–281.

Hao, K. (2019) 'In 2020 le'ts stop AI ethics-washing and actually do something', *MIT Technology Review*. Available from: https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/ [Accessed 5 March 2020].

Human Rights, Big Data and Technology Project (HRBDT). (n.d.) *Identifying Opportunities and Threats to the Right to Health in a New Data-driven Economy*. [Online] Available from: https://www.hrbdt.ac.uk/health/ [Accessed 20 June 2020].

IBM (2019) *Everyday Ethics for Artificial Intelligence*. Available from: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf [Accessed 21 April 2021].

IEEE (2019) *Ethically Aligned Design*. IEEE. Available from: https://ethicsinaction.ieee.org/#ead1e [Accessed 21 April 2021].

Independent High-level Expert Group on Artificial Intelligence (AI HLEG) (2019) *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission.

Internet World Statistics. (2020) *Internet World Stats: Usage and Population Statsistics*. Available from: https://www.internetworldstats.com/stats.htm [Accessed 5 March 2020].

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

Latonero, M. (2018) *Governing Artificial Intelligence: Upholding Huaman Rights and Dignity*. Data & Society. Available from: https://datasociety.net/library/governing-artificial-intelligence/ [Accessed 21 April 2021].

Madiega, T. (2019) *EU Guidelines on Ethics in Artifical Intelligence: Context and Implementation PE 640.163*. Brussels: EPRS | European Parliamentary Research Service.

Merry, S. E. (2006) *Human Rights and Gender Violence: Translating International Law into Local Justice*. Chicago, IL: University of Chicago Press.

Microsoft. (2019). Microsoft Global Human Rights Statement. [Online] Available from: https://www.microsoft.com/en-us/corporate-responsibility/human-rights-statement [Accessed 5 March 2020].

Montreal Declaration Responsible AI (2018) *Montreal Declaration for a Responsible Development of Artificial Intelligence*. Montreal: Université de Montréal and the Fonds de recherche du Québec.

Myers, S., Whittaker, M. and Crawford, K. (2019) *Discriminating Systems: Gender, Race and Power in AI*. New York: AI Now Institute.

Obermeyer, Z., Powers, B., Vogell, C. and Mullainathan, S. (2019) 'Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations', *Science*, 366 (6464), pp. 447–453.

OHCHR. (n.d.) *The Right to Privacy in the Digital Age*. [Online] Available from: https://www.ohchr.org/en/issues/digitalage/pages/digitalageindex.aspx [Accessed 28 June 2020].

OHCHR. (2011) *Guiding Principles on Business and Human Rights*. Geneva: UN Office of the High Commissioner for Human Rights.

OHCHR. (2013) *Who Will be Accountable?*. Geneva: UN Office of the High Commissioner for Human Rights.

Panch, T. (2019) 'Artificial Intelligence: Opportunities and Risks for Public Health', *The Lancet Digital Health*, 1 (1), pp. 13–14.

Rathenau Institute. (2017) *Human Rights in the Robot Age*. Strasbourg: Parliamentary Assembly of the Council of Europe.
Smith, M. J., Axler, R., Bean, S., Rudzicz, F. and Shaw, J. (2020) 'Four equity considerations for the use of artificial intelligence in public health', *Bulletin of WHO*, 98 (4), pp. 290–292.
The Human Rights, Big Data and Technology Project (HRBDT). (2018) *Putting Human Rights at the Heart of the Design, Development and Deployment of Artificial Intelligence*. Colchester: University of Essex, Human Rights Centre.
UN. (1966) *International Covenant on Economic*. Social and Cultural Rights: New York, UN.
UN General Assembly (2018) *Promotion and Protection of the Right to Freedom of Opinion and Expression A/73/348*. New York: UN. Available from: https://undocs.org/pdf?symbol=en/A/73/348 [Accessed 21 April 2021].
UN High Commissioner for Human Rights (2018) *The Right to Privacy in the Digital Age A/HRC/39/29*. Geneva: UN Human Rights Council.
UN High Commissioner for Human Rights (2020) *Impact of new technologies on the promotion and protection of human rights in the context of assemblies*, [Online]. Available from: https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session44/Pages/ListReports.aspx [Accessed 5 July 2020].
UN Secretary General's High-level Panel on Digital Cooperation (HLPDC). (2019) *The Age of Digital Interdependence*. New York: United Nations.
UNI Global Union. (2018) *Top Ten Principles for Ethical Artificial Intelligence*. Nyon: UNI Global Union.
WHO. (2019) *Draft Global Strategy on Digital Health 2020–2024*. Geneva: WHO.
Zuboff, S. (2019) *The Age of Surveillance Capitalism*. New York: Hachette.

Author Information

Sakiko Fukuda-Parr is Professor of International Affairs at The New School in New York. She is also Director of the Independent Panel on Global Governance for Health and Co-Director for The Collective for the Political Determinants of Health at University of Oslo Centre for Development (SUM).

Elizabeth Gibbons is Senior Fellow at the FXB Center for Health and Human Rights in the Harvard T.H. Chan School of Public Health, where she participates in initiatives which leverage her expertise in advancing the human rights of children and adolescents.

Appendix 1 Guidelines analyzed

Public national and multilateral

1. Smart Dubai: *AI Ethics, Principles and Guidelines* (Smart Dubai, 2019)
2. European Parliament: *EU guidelines on ethics in artificial intelligence: Context and implementation* (Tambiama Madiega, 2019)
3. European Commission Independent High-level Expert Group on Artificial Intelligence (AI HLEG): *Ethics Guidelines for Trustworthy AI* (High-level Expert Group on Artificial Intelligence (AI HLEG), 2019)
4. UN Secretary General's High-level Panel on Digital Cooperation: *The Age of Digital Interdependence* (UN Secretary General's High-level Panel on Digital Cooperation (HLPDC), 2019)
5. WHO: Draft Global Strategy on Digital Health 2020-2024 (WHO, 2019)

Multistakeholder

6. IEEE: *Ethically Aligned Design, 1st Edition* (IEEE, March, 2019)
7. *Montreal Declaration for a Responsible Development of Artificial Intelligence* (Montreal Declaration Responsible AI, 2018)

Private Sector

8. IBM: *Everyday Ethics for Artificial Intelligence* (IBM, 2019)
9. Google: *Artificial Intelligence at Google: Our Principles* (Google, 2019)
10. Microsoft: *Global Human Rights Statement* (MICROSOFT, 2019); *Responsible AI* https://www.microsoft.com/en-us/ai/responsible-ai

Civil Society

11. *Amnesty International and Access Now: *The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning* (Amnesty International and Access Now, May 2018)
12. *Access Now: *Human Rights in the Age of Artificial Intelligence* (Access Now, 2018)
13. *UNI Global Union *Top Ten Principles for Ethical Artificial Intelligence* (UNI Global Union, 2018)

Professional Associations

14. American Medical Association, *Policy Recommendations on Augmented Intelligence in Health Care* (AMA, 2018b)
15. American College of Radiology and others, *Ethics of AI in Radiology: European and North American Multi-society Statement*, 2019. (American College of Radiology and others, 2019)