nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0088-2

The global landscape of AI ethics guidelines

Anna Jobin, Marcello Ienca and Effy Vayena*

In the past five years, private companies, research institutions and public sector organizations have issued principles and guidelines for ethical artificial intelligence (AI). However, despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. To investigate whether a global agreement on these questions is emerging, we mapped and analysed the current corpus of principles and guidelines on ethical AI. Our results reveal a global convergence emerging around five ethical principles (transparency, justice and fairness, non-maleficence, responsibility and privacy), with substantive divergence in relation to how these principles are interpreted, why they are deemed important, what issue, domain or actors they pertain to, and how they should be implemented. Our findings highlight the importance of integrating guideline-development efforts with substantive ethical analysis and adequate implementation strategies.

Artificial intelligence (AI), or the theory and development of computer systems able to perform tasks normally requiring human intelligence, is widely heralded as an ongoing "revolution" transforming science and society altogether[1,2]. While approaches to AI such as machine learning, deep learning and artificial neural networks are reshaping data processing and analysis[3], autonomous and semi-autonomous systems are being increasingly used in a variety of sectors including healthcare, transportation and the production chain[4]. In light of its powerful transformative force and profound impact across various societal domains, AI has sparked ample debate about the principles and values that should guide its development and use[5,6]. Fears that AI might jeopardize jobs for human workers[7], be misused by malevolent actors[8], elude accountability or inadvertently disseminate bias and thereby undermine fairness[9] have been at the forefront of the recent scientific literature and media coverage. Several studies have discussed the topic of ethical AI[10–13], notably in meta-assessments[14–16] or in relation to systemic risks[17,18] and unintended negative consequences such as algorithmic bias or discrimination[19–21].

National and international organizations have responded to these concerns by developing ad hoc expert committees on AI, often mandated to draft policy documents. These committees include the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organisation for Economic Co-operation and Development (OECD), the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, and the Select Committee on Artificial Intelligence of the UK House of Lords. As part of their institutional appointments, these committees have produced or are reportedly producing reports and guidance documents on AI. Similar efforts are taking place in the private sector, especially among corporations who rely on AI for their business. In 2018 alone, companies such as Google and SAP publicly released AI guidelines and principles. Declarations and recommendations have also been issued by professional associations and non-profit organizations such as the Association of Computing Machinery (ACM), Access Now and Amnesty International. This proliferation of soft-law efforts can be interpreted as a governance response to advanced research into AI, whose research output and market size have drastically increased[22] in recent years.

Reports and guidance documents for ethical AI are instances of what is termed non-legislative policy instruments or soft law[23]. Unlike so-called hard law—that is, legally binding regulations passed by the legislatures to define permitted or prohibited conduct—ethics guidelines are not legally binding but persuasive in nature. Such documents are aimed at assisting with—and have been observed to have significant practical influence on—decision-making in certain fields, comparable to that of legislative norms[24]. Indeed, the intense efforts of such a diverse set of stakeholders in issuing AI principles and policies is noteworthy, because they demonstrate not only the need for ethical guidance, but also the strong interest of these stakeholders to shape the ethics of AI in ways that meet their respective priorities[16,25]. Specifically, the private sector's involvement in the AI ethics arena has been called into question for potentially using such high-level soft policy as a portmanteau to either render a social problem technical[16] or to eschew regulation altogether[26]. Beyond the composition of the groups that have produced ethical guidance on AI, the content of this guidance itself is of interest. Are these various groups converging on what ethical AI should be, and the ethical principles that will determine the development of AI? If they diverge, what are their differences and can these differences be reconciled?

Our Perspective maps the global landscape of existing ethics guidelines for AI and analyses whether a global convergence is emerging regarding both the principles for ethical AI and the suggestions regarding its realization. This analysis will inform scientists, research institutions, funding agencies, governmental and intergovernmental organizations, and other relevant stakeholders involved in the advancement of ethically responsible innovation in AI.

Methods

We conducted a scoping review of the existing corpus of documents containing soft-law or non-legal norms issued by organizations. This included a search for grey literature containing principles and guidelines for ethical AI, with academic and legal sources excluded. A scoping review is a method aimed at synthesizing and mapping the existing literature[27] that is considered particularly suitable for complex or heterogeneous areas of research[27,28]. Given the absence of a unified database for AI-specific ethics guidelines, we developed a protocol for discovery and eligibility, adapted from the Preferred

Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. *e-mail: effy.vayena@hest.ethz.ch

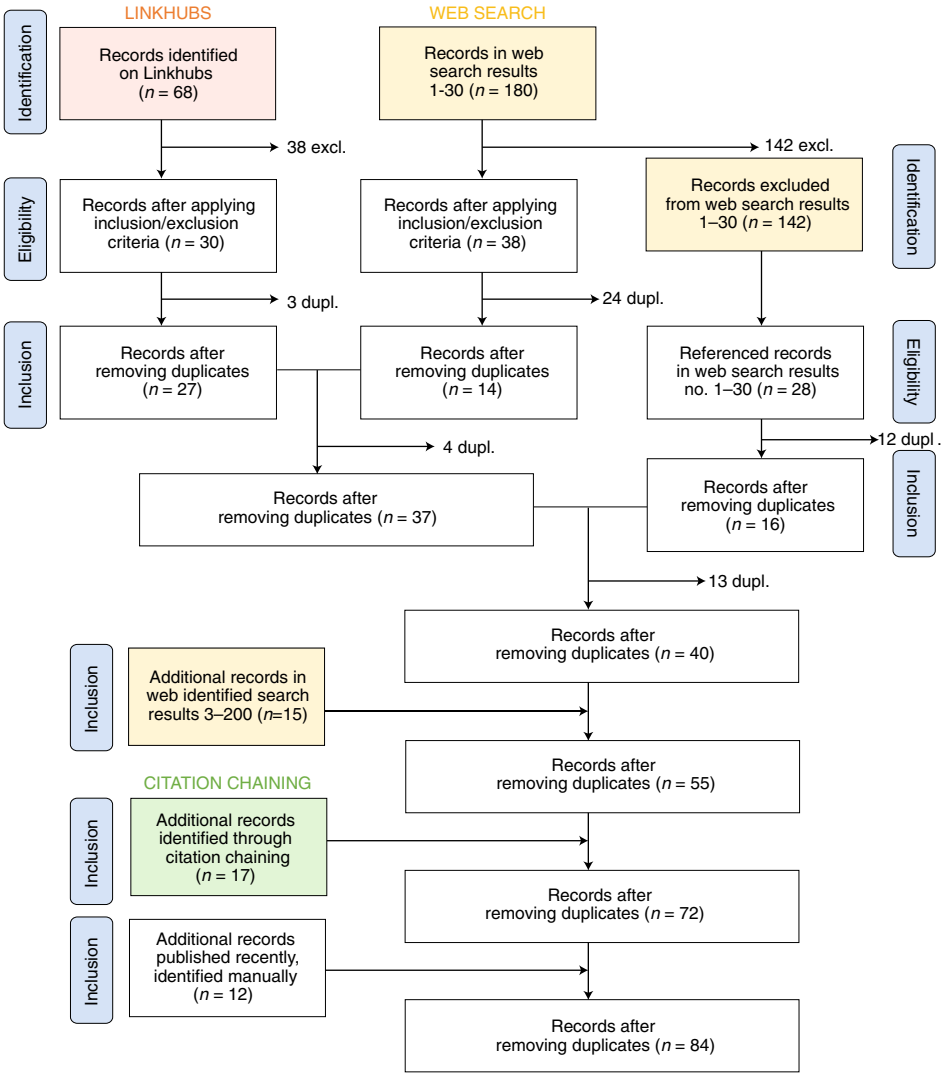

Fig. 1 | PRISMA-based flowchart of retrieval process. Flowchart of our retrieval process based on the PRISMA template for systematic reviews[36]. We relied on three search strategies (linkhubs, web search and citation chaining) and added the most recent records manually, identifying a total of 84 eligible, non-duplicate documents containing ethical principles for AI.

Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework[29]. The protocol was pilot-tested and calibrated prior to data collection. Following best practices for grey literature retrieval, a multi-stage screening strategy involving both inductive screening via search engine and deductive identification of relevant entities with associated websites and online collections was conducted. To achieve comprehensiveness and systematicity, relevant documents were retrieved by relying on three sequential search strategies (Fig. 1): first, a manual search of four link hub webpages ('linkhubs')[30–33] was performed. Sixty-eight sources were retrieved, out of which 30 were eligible (27 after removing duplicates). Second, a keyword-based web search of the Google.com search engine was performed in private-browsing mode, after logging out from personal accounts and erasing all web cookies and history[34,35]. The search was performed using the following keywords: 'AI principles', 'artificial intelligence principles', 'AI guidelines', 'artificial intelligence guidelines', 'ethical AI' and 'ethical artificial intelligence'. Every link in the first 30 search results was followed and screened (1) for AI principles, resulting in ten more sources after removing duplicates, and (2) for articles mentioning AI principles, leading to the identification of three additional non-duplicate sources. The remaining Google results up to the 200th listings for each Google search were followed and screened for AI principles only. Within these additional 1,020 link listings we identified 15 non-duplicate documents. After identifying relevant documents through the two processes described, we used citation chaining to manually screen the full texts and, if applicable, reference lists of all eligible sources in order to identify other relevant documents. Seventeen additional sources were identified. We continued to monitor the literature in parallel with the data analysis and until 23 April 2019 to retrieve eligible documents that were released after our search was completed. Twelve new sources were included within this extended time frame. To ensure theoretical saturation, we exhausted the citation chaining within all identified sources until no additional relevant document could be identified.

Based on our inclusion/exclusion criteria, policy documents (including principles, guidelines and institutional reports) included in the final synthesis were (1) written in English, German, French, Italian or Greek; (2) issued by institutional entities from both the private and the public sectors; (3) referred explicitly in their title/description to AI or ancillary notions; and (4) expressed a normative ethical stance defined as a moral preference for a defined course of action (Supplementary Table 1). Following full-text screening, 84 sources or parts thereof were included in the final synthesis (Supplementary Table 2).

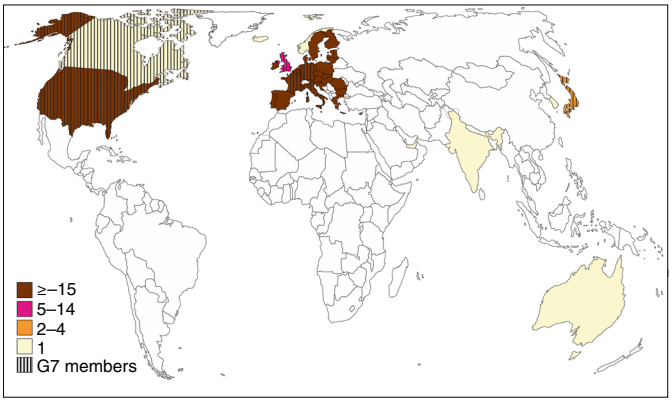

Fig. 2 | Geographic distribution of issuers of ethical AI guidelines by number of documents released. Most ethics guidelines are released in the United States ($n = 21$) and within the European Union (19), followed by the United Kingdom (13) and Japan (4). Canada, Iceland, Norway, the United Arab Emirates, India, Singapore, South Korea and Australia are represented with 1 document each. Having endorsed a distinct G7 statement, member states of the G7 countries are highlighted separately. Map created using https://d-maps.com/carte.php?num_car=13181.

Content analysis of the 84 sources was independently conducted by two researchers in two cycles of manual coding and one cycle of code mapping within the qualitative data analysis software NVivo for Mac version 11.4. During the first cycle of coding, one researcher exhaustively tagged all relevant text through inductive coding[37], attributing a total of 3,457 codes, out of which 1,180 were subsequently discovered to pertain to ethical principles. Subsequently, two researchers conducted the code-mapping process in order to reduce subjective bias. The process of code mapping, a method for qualitative metasynthesis[38], consisted of two iterations of theming[37], whereby categories were first attributed to each code, then categorized in turn (Supplementary Table 3). For the theming of ethical principles, we relied deductively on normative ethical literature. Ethical categories were inspected and assessed for consistency by two researchers with primary expertise in ethics. Thirteen ethical categories emerged from code mapping, two of which were merged with others due to independently assessed semantic and thematic proximity. Finally, we extracted significance and frequency by applying focused coding, a second cycle coding methodology used for interpretive analysis[37], to the data categorized in ethical categories. A consistency check was performed both by reference to the relevant ethics literature and a process of deliberative mutual adjustment among the general principles and the particular judgments contained in the policy documents, an analytic strategy known as reflective equilibrium[39].

Results

Our search identified 84 documents containing ethical principles or guidelines for AI (Tables 1 and 2). Data reveal a significant increase over time in the number of publications, with 88% having been released after 2016 (Supplementary Table 2). Data breakdown by type and geographic location of issuing organization (Supplementary Table 2) shows that most documents were produced by private companies ($n = 19$; 22.6%) and governmental agencies respectively ($n = 18$; 21.4%), followed by academic and research institutions ($n = 9$; 10.7%), intergovernmental or supranational organizations ($n = 8$; 9.5%), non-profit organizations and professional associations/scientific societies ($n = 7$ each; 8.3% each), private sector alliances ($n = 4$; 4.8%), research alliances ($n = 1$; 1.2%), science foundations ($n = 1$; 1.2%), federations of worker unions ($n = 1$; 1.2%) and political parties ($n = 1$; 1.2%). Four documents were issued by initiatives belonging to more than one of the above categories and four more could not be classified at all (4.8% each).

In terms of geographic distribution, data show a prominent representation of more economically developed countries, with the USA ($n = 21$; 25%) and the UK ($n = 13$; 15.5%) together accounting for more than a third of all ethical AI principles, followed by Japan ($n = 4$; 4.8%), Germany, France and Finland (each $n = 3$; 3.6% each). The cumulative number of sources from the European Union, comprising both documents issued by EU institutions ($n = 6$) and documents issued within each member state (13 in total), accounts for 19 documents overall. African and South-American countries are not represented independently from international or supranational organizations (Fig. 2).

Data breakdown by target audience indicates that most principles and guidelines are addressed to multiple stakeholder groups ($n = 27$; 32.1%). Another significant portion of the documents is self-directed, as they are addressed to a category of stakeholders within the sphere of activity of the issuer such as the members of the issuing organization or the issuing company's employees ($n = 24$; 28.6%). Finally, some documents target the public sector ($n = 10$; 11.9%), the private sector ($n = 5$; 6.0%), or other specific stakeholders beyond members of the issuing organization, namely developers or designers ($n = 3$; 3.6%), "organizations" ($n = 1$; 1.2%) and researchers ($n = 1$; 1.2%). Thirteen sources (15.5%) do not specify their target audience (Supplementary Table 1).

Eleven overarching ethical values and principles have emerged from our content analysis. These are, by frequency of the number of sources in which they were featured: transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity (Table 3).

No single ethical principle appeared to be common to the entire corpus of documents, although there is an emerging convergence around the following principles: transparency, justice and fairness, non-maleficence, responsibility and privacy. These principles are referenced in more than half of all the sources. Nonetheless, further thematic analysis reveals significant semantic and conceptual divergences in both how the 11 ethical principles are interpreted and the specific recommendations or areas of concern derived from each. A detailed thematic evaluation is presented in the following.

Transparency. Featured in 73 of our 84 sources, transparency is the most prevalent principle in the current literature. Thematic analysis reveals significant variation in relation to the interpretation, justification, domain of application and mode of achievement. References to transparency comprise efforts to increase explainability, interpretability or other acts of communication and disclosure (Table 3). Principal domains of application include data use[40–43], human–AI interaction[40,44–52], automated decisions[43,53–63] and the purpose of data use or application of AI systems[41,44,64–68]. Primarily, transparency is presented as a way to minimize harm and improve AI[53–55,61,62,66,69–72], though some sources underline its benefit for legal reasons[54,62,63,66,67,69] or to foster trust[40,41,46,50,53,54,65,68,69,73–75]. A few sources also link transparency to dialogue, participation and the principles of democracy[47,58,66,67,69,76].

To achieve greater transparency, many sources suggest increased disclosure of information by those developing or deploying AI systems[53,68,77,78], although specifications regarding what should be communicated vary greatly: use of AI[62], source code[48,69,79], data use[52,64,67,75], evidence base for AI use[74], limitations[42,50,64,68,75,77,80], laws[79,81], responsibility for AI[57], investments in AI[61,82] and possible impact[83]. The provision of explanations "in non-technical terms"[43] or auditable by humans[54,77] is encouraged. Whereas audits and auditability[45,56,61,62,67,76,78,79,84,85] are mainly proposed by data protection offices and non-profit organizations, it is mostly the private sector that suggests technical solutions[44,47,69,76,86,87]. Alternative measures focus on oversight[62,64,65,72,79], interaction and mediation

Table 1 | Ethics guidelines for AI by country of issuer (Australia–UK)

Name of document/website	Issuer	Country of issuer
Artificial Intelligence. Australia's Ethics Framework: A Discussion Paper	Department of Industry Innovation and Science	Australia
Montréal Declaration: Responsible AI	Université de Montréal	Canada
Work in the Age of Artificial Intelligence. Four Perspectives on the Economy, Employment, Skills and Ethics	Ministry of Economic Affairs and Employment	Finland
Tieto's AI Ethics Guidelines	Tieto	Finland
Commitments and Principles	OP Group	Finland
How Can Humans Keep the Upper Hand? Report on the Ethical Matters Raised by AI Algorithms	French Data Protection Authority (CNIL)	France
For a Meaningful Artificial Intelligence. Towards a French and European Strategy	Mission Villani	France
Ethique de la Recherche en Robotique	CERNA (Allistene)	France
AI Guidelines	Deutsche Telekom	Germany
SAP's Guiding Principles for Artificial Intelligence	SAP	Germany
Automated and Connected Driving: Report	Federal Ministry of Transport and Digital Infrastructure, Ethics Commission	Germany
Ethics Policy	Icelandic Institute for Intelligent Machines (IIIM)	Iceland
Discussion Paper: National Strategy for Artificial Intelligence	National Institution for Transforming India (NITI Aayog)	India
L'intelligenzia Artificiale al Servizio del Cittadino	Agenzia per l'Italia Digitale (AGID)	Italy
The Japanese Society for Artificial Intelligence Ethical Guidelines	Japanese Society for Artificial Intelligence	Japan
Report on Artificial Intelligence and Human Society (unofficial translation)	Advisory Board on Artificial Intelligence and Human Society (initiative of the Minister of State for Science and Technology Policy)	Japan
Draft AI R&D Guidelines for International Discussions	Institute for Information and Communications Policy (IICP), The Conference toward AI Network Society	Japan
Sony Group AI Ethics Guidelines	Sony	Japan
Human Rights in the Robot Age Report	The Rathenau Institute	Netherlands
Dutch Artificial Intelligence Manifesto	Special Interest Group on Artificial Intelligence (SIGAI), ICT Platform Netherlands (IPN)	Netherlands
Artificial Intelligence and Privacy	The Norwegian Data Protection Authority	Norway
Discussion Paper on Artificial Intelligence (AI) and Personal Data—Fostering Responsible Development and Adoption of AI	Personal Data Protection Commission Singapore	Singapore
Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society	Government of the Republic of Korea	South Korea
AI Principles of Telefónica	Telefónica	Spain
AI Principles & Ethics	Smart Dubai	UAE
Principles of robotics	Engineering and Physical Sciences Research Council UK (EPSRC)	UK
The Ethics of Code: Developing AI for Business with Five Core Principles	Sage	UK
Big Data, Artificial Intelligence, Machine Learning and Data Protection	Information Commissioner's Office	UK
DeepMind Ethics & Society Principles	DeepMind Ethics & Society	UK
Business Ethics and Artificial Intelligence	Institute of Business Ethics	UK
AI in the UK: Ready, Willing and Able?	UK House of Lords, Select Committee on Artificial Intelligence	UK
Artificial Intelligence (AI) in Health	Royal College of Physicians	UK
Initial Code of Conduct for Data-Driven Health and Care Technology	UK Department of Health & Social Care	UK
Ethics Framework: Responsible AI	Machine Intelligence Garage Ethics Committee	UK
The Responsible AI Framework	PricewaterhouseCoopers UK	UK
Responsible AI and Robotics. An Ethical Framework.	Accenture UK	UK
Machine Learning: The Power and Promise of Computers that Learn by Example	The Royal Society	UK
Ethical, Social, and Political Challenges of Artificial Intelligence in Health	Future Advocacy	UK

Table 2 | Ethics guidelines for AI by country of issuer (USA, international, EU and N/A)

Name of document/website	Issuer	Country of issuer
Unified Ethical Frame for Big Data Analysis. IAF Big Data Ethics Initiative, Part A	The Information Accountability Foundation	USA
The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA
Statement on Algorithmic Transparency and Accountability	Association for Computing Machinery (ACM)	USA
AI Principles	Future of Life Institute	USA
AI—Our Approach	Microsoft	USA
Artificial Intelligence. The Public Policy Opportunity	Intel Corporation	USA
IBM's Principles for Trust and Transparency	IBM	USA
OpenAI Charter	OpenAI	USA
Our Principles	Google	USA
Policy Recommendations on Augmented Intelligence in Health Care H-480.940	American Medical Association (AMA)	USA
Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers and Developers	IBM	USA
Governing Artificial Intelligence. Upholding Human Rights & Dignity	Data & Society	USA
Intel's AI Privacy Policy White Paper. Protecting Individuals' Privacy and Data in the Artificial Intelligence World	Intel Corporation	USA
Introducing Unity's Guiding Principles for Ethical AI—Unity Blog	Unity Technologies	USA
Digital Decisions	Center for Democracy & Technology	USA
Science, Law and Society (SLS) Initiative	The Future Society	USA
AI Now 2018 Report	AI Now Institute	USA
Responsible Bots: 10 Guidelines for Developers of Conversational AI	Microsoft	USA
Preparing for the Future of Artificial Intelligence	Executive Office of the President; National Science and Technology Council; Committee on Technology	USA
The National Artificial Intelligence Research and Development Strategic Plan	National Science and Technology Council; Networking and Information Technology Research and Development Subcommittee	USA
AI Now 2017 Report	AI Now Institute	USA
Position on Robotics and Artificial Intelligence	The Greens (Green Working Group Robots)	EU
Report with Recommendations to the Commission on Civil Law Rules on Robotics	European Parliament	EU
Ethics Guidelines for Trustworthy AI	High-Level Expert Group on Artificial Intelligence	EU
AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	AI4People	EU
European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment	Council of Europe: European Commission for the Efficiency of Justice (CEPEJ)	EU
Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems	European Commission, European Group on Ethics in Science and New Technologies	EU
Artificial Intelligence and Machine Learning: Policy Paper	Internet Society	International
Report of COMEST on Robotics Ethics	COMEST/UNESCO	International
Ethical Principles for Artificial Intelligence and Data Analytics	Software & Information Industry Association (SIIA), Public Policy Division	International
ITI AI Policy Principles	Information Technology Industry Council (ITI)	International
Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	International
Top 10 Principles for Ethical Artificial Intelligence	UNI Global Union	International
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI	International
White Paper: How to Prevent Discriminatory Outcomes in Machine Learning	WEF, Global Future Council on Human Rights 2016-2018	International

Continued

Table 2 | Ethics guidelines for AI by country of issuer (USA, international, EU and N/A) (Continued)

Name of document/website	Issuer	Country of issuer
Privacy and Freedom of Expression in the Age of Artificial Intelligence	Privacy International & Article 19	International
The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems	Access Now; Amnesty International	International
Charlevoix Common Vision for the Future of Artificial Intelligence	Leaders of the G7	International
Artificial Intelligence: Open Questions About Gender Inclusion	W20	International
Declaration on Ethics and Data Protection in Artificial Intelligence	ICDPPC	International
Universal Guidelines for Artificial Intelligence	The Public Voice	International
Ethics of AI in Radiology: European and North American Multisociety Statement	American College of Radiology; European Society of Radiology; Radiology Society of North America; Society for Imaging Informatics in Medicine; European Society of Medical Imaging Informatics; Canadian Association of Radiologists; American Association of Physicists in Medicine	International
Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition (EAD1e)	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	International
Tenets	Partnership on AI	N/A
Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Fairness, Accountability, and Transparency in Machine Learning (FATML)	N/A
10 Principles of Responsible AI	Women Leading in AI	N/A

with stakeholders and the public[41,49,53,68,78,88] and the facilitation of whistleblowing[53,77].

Justice, fairness and equity. Justice is mainly expressed in terms of fairness[40,42,44–46,65,67,75,77,83,89–94], and of prevention, monitoring or mitigation of unwanted bias[40,45,50,57,64,69,71,75,81,86,90,91,95–97] and discrimination[45,50,53,55,61,62,67,72,73,77,85,98–101], the latter being significantly less referenced than the first two by the private sector. Whereas some sources focus on justice as respect for diversity[48,55,73,76,82,83,87,89,95,97,102,103], inclusion[48,62,64,68,89,97] and equality[58,62,68,76,77,89,95], others call for a possibility to appeal or challenge decisions[45,52–54,91,96], or the right to redress[50,59,62,63,67,85,102] and remedy[62,65]. Sources also emphasize the importance of fair access to AI[76,87,104], data[50,54,61,84,100,105–107] and the benefits of AI[54,55,97,108]. Issuers from the public sector place particular emphasis on AI's impact on the labour market[54,55,72,101,109], and the need to address democratic[50,55,76,90] or societal[48,65,72,82] issues. Sources focusing on the risk of biases within datasets underline the importance of acquiring and processing accurate, complete and diverse data[40,45,69,87,110], especially training data[44,50,52,55,69,75].

If specified, the preservation and promotion of justice are proposed to be pursued through: (1) technical solutions such as standards[67,85,106] or explicit normative encoding[45,54,60,84]; (2) transparency[71,79], notably by providing information[53,55,96] and raising public awareness of existing rights and regulation[45,76]; (3) testing[69,75,84,86], monitoring[71,73] and auditing[56,63,67,84], the preferred solution of notably data protection offices; (4) developing or strengthening the rule of law and the right to appeal, recourse, redress or remedy[54,55,59,62,63,65,85,91,96]; and (5) via systemic changes and processes such as governmental action[59,62,104,109] and oversight[111], a more interdisciplinary[64,82,102,110] or otherwise diverse[75,76,87,102,104,112] workforce, as well as better inclusion of civil society or other relevant stakeholders in an interactive manner[45,50,58,63,72,74,75,82,85,86,96,97,103] and increased attention to the distribution of benefits[42,50,55,65,80,93].

Non-maleficence. References to non-maleficence occur significantly more often than references to beneficence and encompass general calls for safety and security[97,107,113,114] or state that AI should never cause foreseeable or unintentional harm[40,47,50,73,77,96]. More granular considerations entail the avoidance of specific risks or potential harms—for example, intentional misuse via cyberwarfare and malicious hacking[68,70,71,95,98,106]—and suggest risk-management strategies. Harm is primarily interpreted as discrimination[55,61,64,65,67,112,115], violation of privacy[40,52,61,81,95,115,116] or bodily harm[42,47,48,50,73,109,113,117]. Less frequent characterizations include loss of trust[47] or skills[61]; "radical individualism"[55]; the risk that technological progress might outpace regulatory measures[74]; and negative impacts on long-term social well-being[61], infrastructure[61], or psychological[52,73], emotional[73] or economic aspects[61,73].

Harm-prevention guidelines focus primarily on technical measures and governance strategies, ranging from interventions at the level of AI research[44,64,81,96,102,118], design[40,42,44,49,56,73,75], technology development and/or deployment[71] to lateral and continuous approaches[50,72,80]. Technical solutions include in-built data quality evaluations[42] or security[40] and privacy by design[40,44,56], though notable exceptions also advocate for establishing industry standards[47,81,119]. Proposed governance strategies include active cooperation across disciplines and stakeholders[50,64,70,79], compliance with existing or new legislation[44,48,52,98,112,116], and the need to establish oversight processes and practices—notably tests[53,55,64,91,96], monitoring[53,75], audits and assessments by internal units, customers, users, independent third parties or governmental entities[57,65,68,75,98,111,112,115], often geared towards standards for AI implementation and outcome assessment. Many imply that damages may be unavoidable, in which case risks should be assessed[57,65,68], reduced[57,86,89–91] and mitigated[51,52,55,70,80,85], and the attribution of liability should be clearly defined[48,54,55,61,99]. Several sources mention potential "multiple"[45,48,49,69,77,96] or "dual-use"[8,50,55], take explicit position against military application[48,54,117] or simply guard against the dynamics of an "arms race"[51,70,119].

Responsibility and accountability. Despite widespread references to "responsible AI"[60,68,95,100], responsibility and accountability are rarely defined. Nonetheless, specific recommendations include acting with "integrity"[64,69,77] and clarifying the attribution of responsibility and legal liability[40,75,95,120], if possible upfront[53], in contracts[69] or, alternatively, by centring on remedy[43]. In contrast, other sources suggest focusing on the underlying reasons and processes that may

Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

lead to potential harm[91,100]. Yet others underline the responsibility of whistleblowing in case of potential harm[53,72,77], and aim at promoting diversity[66,109] or introducing ethics into science, technology, engineering and mathematics education[76]. Very different actors are named as being responsible and accountable for AI's actions and decisions: AI developers[75,77,90,113], designers[53,61], "institutions"[57,59] or "industry"[86]. Further divergence emerged on whether AI should be held accountable in a human-like manner[87] or whether humans should always be the only actors who are ultimately responsible for technological artifacts[48,49,52,54,69,109].

Privacy. Ethical AI sees privacy both as a value to uphold[61,81,92,116] and as a right to be protected[44,45,54,55,70]. While often undefined, privacy is frequently presented in relation to data protection[40,44,53,70,75,83,88,96,100,115] and data security[44,52,81,83,105,115]. A few sources link privacy to freedom[55,70] or trust[91,109]. Suggested modes of achievement fall into three categories: technical solutions[81,97], such as differential privacy[91,106], privacy by design[42,44,45,96,115], data minimization[53,75] and access control[53,75]; calls for more research[64,81,91,115] and awareness[81,91]; and regulatory approaches[42,69,88], with sources referring to legal compliance more broadly[44,49,53,75,77,98], or suggesting certificates[121] or the creation or adaptation of laws and regulations to accommodate the specificities of AI[81,91,105,122].

Beneficence. While promoting good ('beneficence' in ethical terms) is often mentioned, it is rarely defined, though notable exceptions mention the augmentation of human senses[103], the promotion of human well-being and flourishing[51,107], peace and happiness[77], the creation of socio-economic opportunities[53], and economic prosperity[54,70]. Similar uncertainty concerns the actors that should benefit from AI: private sector issuers tend to highlight the benefit of AI for customers[40,65], though overall many sources require AI to be shared[66,69,93] and to benefit everyone[53,76,82,101], "humanity"[54,61,77,117,119], both of the above[65,83], "society"[51,104], "as many people as possible"[54,70,116], "all sentient creatures"[100], the "planet"[54,89] and the environment[55,107]. Strategies for the promotion of good include aligning AI with human values[51,61], advancing "scientific understanding of the world"[117], minimizing power concentration[119] or, conversely, using power "for the benefit of human rights"[99], working more closely with "affected" people[82], minimizing conflicts of interests[119], proving beneficence through customer demand[65] and feedback[75], and developing new metrics and measurements for human well-being[61,107].

Freedom and autonomy. Whereas some sources specifically refer to the freedom of expression[45,90,99,122] or informational self-determination[45,107] and "privacy-protecting user controls"[75], others generally promote freedom[48,86,89], empowerment[45,69,116] or autonomy[48,50,79,94,98,113]. Some documents refer to autonomy as a positive freedom, specifically the freedom to flourish[53], to self-determination through democratic means[55], the right to establish and develop relationships with other human beings[55,109], the freedom to withdraw consent[84], or the freedom to use a preferred platform or technology[90,97]. Other documents focus on negative freedom—for example, freedom from technological experimentation[99], manipulation[50] or surveillance[55]. Freedom and autonomy are believed to be promoted through transparency and predictable AI[55], by not "reducing options for and knowledge of citizens"[55], by actively increasing people's knowledge about AI[53,69,79], giving notice and consent[96] or, conversely, by actively refraining from collecting and spreading data in absence of informed consent[47,55,61,72,91].

Trust. References to trust include calls for trustworthy AI research and technology[67,114,116], trustworthy AI developers and organizations[68,77,83], trustworthy "design principles"[108], or underline the importance of customers' trust[40,69,75,83,91,97]. Calls for trust are proposed because a culture of trust among scientists and engineers is believed to support the achievement of other organizational goals[116], or because overall trust in the recommendations, judgments and uses of AI is indispensable for AI to "fulfil its world changing potential"[41]. This last point is contradicted by one guideline explicitly warning against excessive trust in AI[98]. Suggestions for building or sustaining trust include education[50], reliability[67,68], accountability[73], processes to monitor and evaluate the integrity of AI systems over time[68], and tools and techniques ensuring compliance with norms and standards[60,80]. Whereas some guidelines require AI to be transparent[54,60,74,75], understandable[53,54] or explainable[69] in order to build trust, another one explicitly suggests that, instead of demanding understandability, it should be ensured that AI fulfils public expectations[67]. Other reported facilitators of trust include "a Certificate of Fairness"[121], multi-stakeholder dialogue[81], awareness about the value of using personal data[91], and avoiding harm[47,73].

Sustainability. To the extent that is referenced, sustainability calls for development and deployment of AI to consider protecting the environment[50,55,63], improving the planet's ecosystem and biodiversity[54], contributing to fairer and more equal societies[82] and promoting peace[83]. Ideally, AI creates sustainable systems[61,93,107] that process data sustainably[60] and whose insights remain valid over time[65]. To achieve this aim, AI should be designed, deployed and managed with care[55] to increase its energy efficiency and minimize its ecological footprint[48]. To make future developments sustainable, corporations are asked to create policies ensuring accountability in the domain of potential job losses[54] and to use challenges as an opportunity for innovation[55].

Dignity. While dignity remains undefined in existing guidelines, save one specification that it is a prerogative of humans but not robots[109], there is frequent reference to what it entails: dignity is intertwined with human rights[118] or otherwise means avoiding harm[48], forced acceptance[48], automated classification[55] and unknown human–AI interaction[55]. It is argued that AI should not diminish[50] or destroy[97], but respect[99], preserve[86] or even increase human dignity[53,54]. Dignity is believed to be preserved if it is respected by AI developers in the first place[113] and promoted through new legislation[55], through governance initiatives[53], or through government-issued technical and methodological guidelines[99].

Solidarity. Solidarity is mostly referenced in relation to the implications of AI for the labour market[121]. Sources call for a strong social safety net[54,101]. They underline the need for redistributing the benefits of AI in order not to threaten social cohesion[66] and respecting potentially vulnerable persons and groups[50]. Lastly, there is a warning of data collection and practices focused on individuals that may undermine solidarity in favour of "radical individualism"[55].

Discussion

The rapid increase in the number and variety of guidance documents attests to growing interest in AI ethics by the international community and across different types of organization. The nearly equivalent proportion of documents issued by the public sector (that is, governmental and intergovernmental organizations) and the private sector (companies and private sector alliances) suggests that AI ethics concerns both public entities and private enterprises. However, the solutions proposed to meet the ethical challenges diverge significantly. Furthermore, the underrepresentation of geographic areas such as Africa, South and Central America and Central Asia indicates that global regions are not participating equally in the AI ethics debate, which reveals a power imbalance in the international discourse. More economically developed countries are shaping this debate more than others, which raises concerns about neglecting local knowledge, cultural pluralism and the demands of global fairness.

Our thematic synthesis uncovers the existence of 11 clusters of ethical principles and reveals an emerging cross-stakeholder convergence on promoting the ethical principles of transparency, justice, non-maleficence, responsibility and privacy, which are referenced in more than half of all guidelines. The potentially pro-ethical nature of transparency, "enabling or impairing other ethical practices or principles"[123], might partly explain its prevalence, whereas the frequent occurrence of calls for privacy, non-maleficence, justice and fairness can be seen as cautioning the global community against potential risks brought by AI. These themes appear to be intertwined with the theme of responsibility, although only a few guidelines emphasize the duty of all stakeholders involved in the development and deployment of AI to act with integrity.

Because references to non-maleficence outnumber those related to beneficence, it appears that issuers of guidelines are preoccupied with the moral obligation to prevent harm. This focus is corroborated by a greater emphasis on preserving privacy, dignity, autonomy and individual freedom in spite of advances in AI rather than on actively promoting these principles[124]. It could be due to the so-called negativity bias—that is, a general cognitive bias to give greater weight to negative entities[125,126]—or as a precautionary measure to ensure that AI developers and deployers are held in check[25].

The references to trust address a critical ethical issue in AI governance; that is, whether it is morally desirable to foster public trust in AI. Whereas several sources, predominantly from the private sector, highlight the importance of fostering trust in AI through educational and awareness-raising activities, others contend that trust in AI risks diminishing scrutiny and may undermine certain societal obligations of AI producers[127]. This later perspective challenges the dominant view that building public trust in AI should be a fundamental requirement for ethical governance[128].

Sustainability, dignity and solidarity are significantly underrepresented compared to other ethical dimensions, which suggests that these issues might be currently flying under the radar of the mainstream AI ethics debate. This underrepresentation is particularly problematic in light of recent evidence that AI requires massive computational resources, which, in turn, require high energy consumption[129]. The environmental impact of AI, however, involves not only the negative effects of high-footprint digital infrastructures, but also the possibility of harnessing AI for the benefit of ecosystems and the entire biosphere. As the humanitarian cost of anthropogenic climate change is rapidly increasing[130], the principles of sustainability and solidarity appear strictly intertwined. In addition, the ethical principle of solidarity is referenced even more rarely, typically in association with the development of inclusive strategies for the prevention of job losses and unfair sharing of burdens. Little attention is devoted to promoting solidarity through the emerging possibility of using AI expertise for solving humanitarian challenges, a mission that is currently being pursued, among others, by companies such as Microsoft[131] and intergovernmental organizations such as the United Nations Office for Project Services (UNOPS)[132] or the World Health Organization (WHO). A better appraisal of currently underrepresented ethical principles is likely to result in a more inclusive AI ethics landscape.

Although numerical data indicate an emerging convergence around the importance of certain ethical principles, an in-depth thematic analysis paints a more complicated picture. Our focused coding reveals substantive divergences among all 11 ethical principles in relation to four major factors: (1) how ethical principles are interpreted; (2) why they are deemed important; (3) what issue, domain or actors they pertain to; and (4) how they should be implemented. These conceptual and procedural divergences reveal uncertainty as to which ethical principles should be prioritized and how conflicts between ethical principles should be resolved, and it may undermine attempts to develop a global agenda for ethical AI. For example, the need for ever-larger, more diverse datasets to 'unbias' AI might conflict with the requirement to give individuals increased control over their data and its use in order to respect their privacy and autonomy. Similar contrasts emerge between avoiding harm at all costs and the perspective of accepting some degree of harm as long as risks and benefits are weighed against each other. Moreover, risk–benefit evaluations are likely to lead to different results depending on whose well-being will be optimized for and by which actors. Such divergences and tensions illustrate a gap at the cross-section of principle formulation and their implementation into practice[133,134].

These findings have implications for public policy, technology governance and research ethics. At the policy level, greater inter-stakeholder cooperation is needed to mutually align different AI ethics agendas and to seek procedural convergence not only on ethical principles but also their implementation. While global consensus might be desirable it should not come at the cost of obliterating cultural and moral pluralism and may require the development of deliberative mechanisms to adjudicate disagreement among stakeholders from different global regions. Such efforts can be mediated and facilitated by intergovernmental organizations, complemented by bottom-up approaches involving all relevant stakeholders[135]. Furthermore, it should be clarified how AI ethics guidelines relate to existing national and international regulation. Translating principles into practice and seeking harmonization between AI ethics codes (soft law) and legislation (hard law) are important next steps for the global community.

At the level of technology governance, harmonization is typically implemented in terms of standardizations such as the Ethically Aligned Designed initiative[136] led by the Institute of Electrical and

Electronics Engineers (IEEE). Although standardization is a necessary step in the right direction, it remains to be seen whether the impact of these non-legal norms will mostly happen on the policy level or if they will also influence individual practitioners and decision-makers. Finally, research ethics mechanisms like independent review boards will be increasingly required to assess the ethical validity of AI applications in scientific research, especially those in the academic domain. However, AI applications by governments or private corporations are unlikely to fall under their oversight, unless significant expansions to the purview of independent review boards are made.

Limitations

This study has several limitations. First, guidelines and soft-law documents are an instance of grey literature, and are therefore not indexed in conventional scholarly databases. As a result, their retrieval is inevitably less replicable and unbiased compared to systematic database searches of peer-reviewed literature. Following best practices for grey literature review, this limitation has been mitigated by developing a discovery and eligibility protocol, which was pilot-tested prior to data collection. Although search results from search engines are personalized, the risk of personalization influencing discovery has been mitigated through the broadness of both the keyword search and the inclusion of results. A language bias may have skewed our corpus towards English results. Our content analysis presents the typical limitations of qualitative analytic methods. Following best practices for content analysis, this limitation has been mitigated by developing an inductive coding strategy, which was conducted independently by two reviewers to minimize subjective bias. Finally, given the rapid pace of publication of AI guidance documents, there is a possibility that new policy documents were published after our search was completed. To minimize this risk, continuous monitoring of the literature was conducted in parallel with the data analysis and until 23 April 2019. In spite of these limitations, our Perspective provides a comprehensive mapping of the current AI ethics landscape and offers a basis for future research on this topic.

Received: 23 May 2019; Accepted: 23 July 2019;
Published online: 02 September 2019

References

1. Harari, Y. N. Reboot for the AI revolution. *Nature* **550**, 324–327 (2017).
2. Appenzeller, T. The AI revolution in science. *Science* https://doi.org/10.1126/science.aan7064 (2017).
3. Jordan, M. I. & Mitchell, T. M. Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
4. Stead, W. W. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA* **320**, 1107–1108 (2018).
5. Vayena, E., Blasimme, A. & Cohen, I. G. Machine learning in medicine: addressing ethical challenges. *PLOS Med.* **15**, e1002689 (2018).
6. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
7. Science must examine the future of work. *Nature* **550**, 301–302 (2017).
8. Brundage, M. et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, OpenAI, 2018).
9. Zou, J. & Schiebinger, L. AI can be sexist and racist — it's time to make it fair. *Nature* **559**, 324–326 (2018).
10. Boddington, P. *Towards a Code of Ethics for Artificial Intelligence* (Springer, 2017).
11. Bostrom, N. & Yudkowsky, E. in *The Cambridge Handbook of Artificial Intelligence* (eds Frankish, K. & Ramsey, W. M.) 316–334 (Cambridge Univ. Press, 2014). https://doi.org/10.1017/CBO9781139046855.020
12. Etzioni, A. & Etzioni, O. AI assisted ethics. *Ethics Inf. Technol.* **18**, 149–156 (2016).
13. Yuste, R. et al. Four ethical priorities for neurotechnologies and AI. *Nature* **551**, 159–163 (2017).
14. Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M. & Floridi, L. Artificial intelligence and the 'good society': the US, EU, and UK approach. *Sci. Eng. Ethics* **24**, 505–528 (2018).
15. Zeng, Y., Lu, E. & Huangfu, C. Linking artificial intelligence principles. Preprint at https://arxiv.org/abs/1812.04814 (2018).
16. Greene, D., Hoffmann, A. L. & Stark, L. Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proc. 52nd Hawaii International Conference on System Sciences* 2122–2131 (2019).
17. Crawford, K. & Calo, R. There is a blind spot in AI research. *Nature* **538**, 311–313 (2016).
18. Altman, M., Wood, A. & Vayena, E. A harm-reduction framework for algorithmic fairness. *IEEE Security Privacy* **16**, 34–45 (2018).
19. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Preprint at https://arxiv.org/abs/1607.06520 (2016).
20. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).
21. Veale, M. & Binns, R. Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. *Big Data Soc*. https://doi.org/10.1177/2053951717743530 (2017).
22. Shoham, Y. et al. *The AI Index 2018 Annual Report* (AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, 2018).
23. Sossin, L. & Smith, C. W. Hard choices and soft law: ethical codes, policy guidelines and the role of the courts in regulating government. *Alberta Law Rev.* **40**, 867–893 (2003).
24. Campbell, A. & Glass, K. C. The legal status of clinical and ethics policies, codes, and guidelines in medical practice and research. *McGill Law J.* **46**, 473–489 (2001).
25. Benkler, Y. Don't let industry write the rules for AI. *Nature* **569**, 161 (2019).
26. Wagner, B. in *Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* (eds Bayamlioglu, E., Baraliuc, I., Janssens, L. A. W. & Hildebrandt, M.) 84–89 (Amsterdam Univ. Press, 2018).
27. Arksey, H. & O'Malley, L. Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* **8**, 19–32 (2005).
28. Pham, M. T. et al. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Res. Synth. Meth.* **5**, 371–385 (2014).
29. Liberati, A. et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLOS Medicine* **6**, e1000100 (2009).
30. Boddington, P. Alphabetical list of resources. *Ethics for Artificial Intelligence* https://www.cs.ox.ac.uk/efai/resources/alphabetical-list-of-resources/ (2018).
31. Winfield, A. A round up of robotics and AI ethics. *Alan Winfield's Web Log* http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html (2017).
32. National and international AI strategies. *Future of Life Institute* https://futureoflife.org/national-international-ai-strategies/ (2018)
33. Summaries of AI policy resources. *Future of Life Institute* https://futureoflife.org/ai-policy-resources/ (2018).
34. Hagstrom, C., Kendall, S. & Cunningham, H. Googling for grey: using Google and Duckduckgo to find grey literature. In *Abstracts of the 23rd Cochrane Colloquium* Vol. 10, LRO 3.6, 40 (Cochrane Database of Systematic Reviews, 2015).
35. Piasecki, J., Waligora, M. & Dranseika, V. Google search as an additional source in systematic reviews. *Sci. Eng. Ethics* **24**, 809–810 (2017).
36. Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G., The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLOS Med.* **6**, e1000097 (2009).
37. Saldaña, J. *The Coding Manual for Qualitative Researchers* (SAGE, 2013).
38. Noblit, G. W. & Hare, R. D. *Meta-Ethnography: Synthesizing Qualitative Studies* (SAGE, 1988).
39. Daniels, N. *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge Univ. Press, 1996).
40. Guidelines for artificial intelligence. *Deutsche Telekom* https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366 (2018).
41. Transparency and trust in the cognitive era. *IBM* https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/ (2017).
42. Initial code of conduct for data-driven health and care technology. *GOV.UK* https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology (2019).
43. Diakopoulos, N. et al. Principles for accountable algorithms and a social impact statement for algorithms. *FATML* http://www.fatml.org/resources/principles-for-accountable-algorithms (2016).
44. AI principles of Telefónica. *Telefónica* https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles (2018).
45. *Declaration on Ethics and Data Protection in Artificial Intelligence* (Commission Nationale de l'Informatique et des Libertés, European Data Protection Supervisor & Garante per la protezione dei dati personali, 2018).
46. *Everyday Ethics for Artificial Intelligence* (IBM, 2018).

47. *Ethics Commission: Automated and Connected Driving* (Federal Ministry of Transport and Digital Infrastructure, 2017).
48. *Position on Robotics and Artificial Intelligence* (Green Digital Working Group, 2016).
49. Principles of robotics. *EPSRC* https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/ (2011).
50. *Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on AI, 2019).
51. Artificial intelligence principles and ethics. *Smart Dubai* http://www.smartdubai.ae/initiatives/ai-principles-ethics (2019).
52. Dawson, D. et al. *Artificial Intelligence: Australia's Ethics Framework* (Australian Government, 2019).
53. Artificial intelligence and machine learning: policy paper. *Internet Society* https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/ (2017).
54. *Top 10 Principles for Ethical AI* (UNI Global, 2017).
55. *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems* (European Group on Ethics in Science and New Technologies, 2018).
56. *Big Data, Artificial Intelligence, Machine Learning and Data Protection*. (ICO, 2017).
57. Universal guidelines for artificial intelligence. *The Public Voice* https://thepublicvoice.org/ai-universal-guidelines/ (2018).
58. Science, law and society (SLS) initiative. *The Future Society* https://web.archive.org/web/20180621203843/http://thefuturesociety.org/science-law-society-sls-initiative/ (2018).
59. *Statement on Algorithmic Transparency and Accountability* (ACM, 2017).
60. *Dutch Artificial Intelligence Manifesto* (Special Interest Group on Artificial Intelligence, 2018).
61. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2* (IEEE, 2017).
62. The Toronto declaration: protecting the right to equality and non-discrimination in machine learning systems. *Human Rights Watch* https://www.hrw.org/news/2018/07/03/toronto-declaration-protecting-rights-equality-and-non-discrimination-machine (2018).
63. Floridi, L. et al. AI4People—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds Mach.* **28**, 589–707 (2018).
64. SAP's guiding principles for artificial intelligence (AI). *SAP* https://www.sap.com/products/leonardo/machine-learning/ai-ethics.html#guiding-principles (2018).
65. *Ethical Principles for Artificial Intelligence and Data Analytics* (SIIA, 2017).
66. Koski, O. & Husso, K. *Work in the Age of Artificial Intelligence* (Ministry of Economic Affairs and Employment, 2018).
67. Digital decisions. *Center for Democracy & Technology* https://cdt.org/issue/privacy-data/digital-decisions/ (2019).
68. Ethics framework. *MI Garage* https://www.migarage.ai/ethics-framework/ (2019).
69. *Business Ethics and Artificial Intelligence* (Institute of Business Ethics, 2018).
70. Asilomar AI Principles. *Future of Life Institute* https://futureoflife.org/ai-principles/ (2017).
71. The responsible AI framework. *PwC* https://www.pwc.co.uk/services/audit-assurance/risk-assurance/services/technology-risk/technology-risk-insights/accelerating-innovation-through-responsible-ai/responsible-ai-framework.html (2019).
72. Whittaker, M. et al. *AI Now Report 2018* (AI Now Institute, 2018).
73. *Discussion Paper on AI and Personal Data — Fostering Responsible Development and Adoption of AI* (Personal Data Protection Commission Singapore, 2018).
74. Artificial intelligence (AI) in health. *RCP London* https://www.rcplondon.ac.uk/projects/outputs/artificial-intelligence-ai-health (2018).
75. Responsible bots: 10 guidelines for developers of conversational AI. *Microsoft* https://www.microsoft.com/en-us/research/publication/responsible-bots/ (2018).
76. Villani, C. *For a Meaningful Artificial Intelligence: Towards a French and European Strategy* (AI for Humanity, 2018).
77. *The Japanese Society for Artificial Intelligence Ethical Guidelines* (Japanese Society for Artificial Intelligence, 2017).
78. Demiaux, V. *How Can Humans Keep the Upper Hand? The Ethical Matters Raised by Algorithms and Artificial Intelligence* (CNIL, 2017).
79. European ethical charter on the use of artificial intelligence in judicial systems and their environment. *Council of Europe* https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment (2019).
80. *Ethics of AI in Radiology: European and North American Multisociety Statement* (American College of Radiology, 2019).
81. *Charlevoix Common Vision for the Future of Artificial Intelligence* (Leaders of the G7, 2018).
82. DeepMind ethics and society principles. *DeepMind* https://deepmind.com/applied/deepmind-ethics-society/principles/ (2017).
83. *Sony Group AI Ethics Guidelines* (Sony, 2018).
84. *Artificial Intelligence and Privacy* (Datatilsynet, 2018).
85. *White Paper: How to Prevent Discriminatory Outcomes in Machine Learning* (WEF, 2018).
86. *ITI AI Policy Principles* (ITI, 2017).
87. *The Ethics of Code: Developing AI for Business with Five Core Principles* (Sage, 2017).
88. Commitments and principles. *OP* https://www.op.fi/op-financial-group/corporate-social-responsibility/commitments-and-principles (2019).
89. *Tieto's AI Ethics Guidelines* (Tieto, 2018).
90. Introducing Unity's Guiding Principles for Ethical AI. *Unity Blog* https://blogs.unity3d.com/2018/11/28/introducing-unitys-guiding-principles-for-ethical-ai/ (2018).
91. *Discussion Paper: National Strategy for Artificial Intelligence* (NITI Aayog, 2018).
92. *AI in the UK: Ready, Willing and Able* 183 (House of Lords, 2018).
93. *Unified Ethical Frame for Big Data Analysis: IAF Big Data Ethics Initiative, Part A* (The Information Accountability Foundation, 2015).
94. Fenech, M., Strukelj, N. & Buston, O. *Ethical, Social, and Political Challenges of Artificial Intelligence in Health* (Future Advocacy, 2019).
95. Responsible AI and robotics: an ethical framework. *Accenture* https://www.accenture.com/gb-en/company-responsible-ai-robotics (2019).
96. Artificial intelligence at Google: our principles. *Google AI* https://ai.google/principles/ (2019).
97. Microsoft AI principles. *Microsoft* https://www.microsoft.com/en-us/ai/our-approach-to-ai (2017).
98. *Éthique de la Recherche en Robotique* (Allistene, 2014).
99. van Est, R. & Gerritsen, J. *Human Rights in the Robot Age: Challenges Arising from the Use of Robotics, Artificial Intelligence, and Virtual and Augmented Reality* (Rathenau Institute, 2017).
100. The declaration. *Montreal Declaration* https://www.montrealdeclaration-responsibleai.com/the-declaration (2017).
101. *Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society: Managing the Fourth Industrial Revolution* (Government of the Republic of Korea, 2017).
102. Crawford, K. et al. *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* (AI Now Institute, 2016).
103. *Report on Artificial Intelligence and Human Society: Unofficial Translation* (Ministry of State for Science and Technology Policy, 2017).
104. *Preparing for the future of Artificial Intelligence* (NSTC, 2016).
105. *Artificial Intelligence: The Public Policy Opportunity* (Intel, 2017).
106. *Machine Learning: The Power and Promise of Computers that Learn by Example* (Royal Society, 2017).
107. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 1* (IEEE, 2019).
108. *Report with Recommendations to the Commission on Civil Law Rules on Robotics* (European Parliament, 2017).
109. *Report of COMEST on Robotics Ethics* (COMEST/UNESCO, 2017).
110. Campolo, A., Sanfilippo, M., Whittaker, M. & Crawford, K. *AI Now 2017 Report* (AI ow Institute, 2017).
111. *Policy Recommendations on Augmented Intelligence in Health Care H-480.940* (AMA, 2018).
112. Avila, R., Brandusescu, A., Freuler, J. O. & Thakur, D. *Artificial Intelligence: Open Questions about Gender Inclusion* (World Wide Web Foundation, 2018).
113. *Draft AI R&D Guidelines for International Discussions* (The Conference toward AI Network Society, 2017).
114. *The National Artificial Intelligence Research and Development Strategic Plan* (NSTC, 2016).
115. Hoffmann, D. & Masucci, R. *Intel's AI Privacy Policy White Paper: Protecting Individuals' Privacy and Data In The Artificial Intelligence World* (Intel, 2018).
116. Tenets. *The Partnership on AI* https://www.partnershiponai.org/tenets/ (2016).
117. Ethics Policy. *IIIM* http://www.iiim.is/2015/08/ethics-policy/ (2015).
118. Latonero, M. Governing artificial intelligence: upholding human rights & dignity. *Data & Society* https://datasociety.net/output/governing-artificial-intelligence/ (2018).
119. OpenAI Charter. *OpenAI* https://blog.openai.com/openai-charter/ (2018).
120. *L'Intelligenzia Artificiale al Servizio del Cittadino* (AGID, 2018).
121. Gilburt, B. Women leading in AI: 10 principles of responsible AI. *Towards Data Science* https://towardsdatascience.com/women-leading-in-ai-10-principles-for-responsible-ai-8a167fc09b7d (2019).
122. *Privacy and Freedom of Expression in the Age of Artificial Intelligence* (Privacy International/Article 19, 2018).
123. Turilli, M. & Floridi, L. The ethics of information transparency. *Ethics Inf. Technol.* **11**, 105–112 (2009).

124. Taddeo, M. & Floridi, L. How AI can be a force for good. *Science* **361**, 751–752 (2018).
125. Rozin, P. & Royzman, E. B. Negativity bias, negativity dominance, and contagion. *Person. Soc. Psychol. Rev.* https://doi.org/10.1207/S15327957PSPR0504_2 (2016).
126. Bentley, P. J., Brundage, M., Häggström, O. & Metzinger, T. *Should We Fear Artificial Intelligence? In-Depth Analysis* (European Parliament, 2018).
127. Bryson, J. AI & global governance: no one should trust AI. *United Nations University* https://cpr.unu.edu/ai-global-governance-no-one-should-trust-ai.html (2018).
128. Winfield, A. F. T. & Marina, J. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philos. Trans. R. Soc. A* **376**, 20180085 (2018).
129. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for deep learning in NLP. Preprint at https://arxiv.org/abs/1906.02243 (2019).
130. Scheffran, J., Brzoska, M., Kominek, J., Link, P. M. & Schilling, J. Climate change and violent conflict. *Science* **336**, 869–871 (2012).
131. AI for humanitarian action. *Microsoft* https://www.microsoft.com/en-us/ai/ai-for-humanitarian-action (2019).
132. Lancaster, C. Can artificial intelligence improve humanitarian responses? *UNOPS* https://www.unops.org/news-and-stories/insights/can-artificial-intelligence-improve-humanitarian-responses (2018).
133. Hagendorff, D. T. The ethics of AI ethics — an evaluation of guidelines. Preprint at https://arxiv.org/abs/1903.03425 (2019).
134. Whittlestone, J., Nyrup, R., Alexandrova, A. & Cave, S. The role and limits of principles in AI ethics: towards a focus on tensions. In *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society* 195–200 (2019).
135. Mittelstadt, B. AI ethics — too principled to fail? Preprint at https://arxiv.org/abs/1906.06668 (2019).
136. The IEEE global initiative on ethics of autonomous and intelligent systems. *IEEE Standards Association* https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (2019).

Acknowledgements

The authors would like to thank J. Sleigh for her help with creating the colour-coded map.

Author contributions

E.V. conceived the research; A.J., M.I. and E.V. designed the research; A.J. performed the research; A.J. and M.I. analysed the data; A.J., M.I. and E.V. wrote the paper.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0088-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to E.V.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.