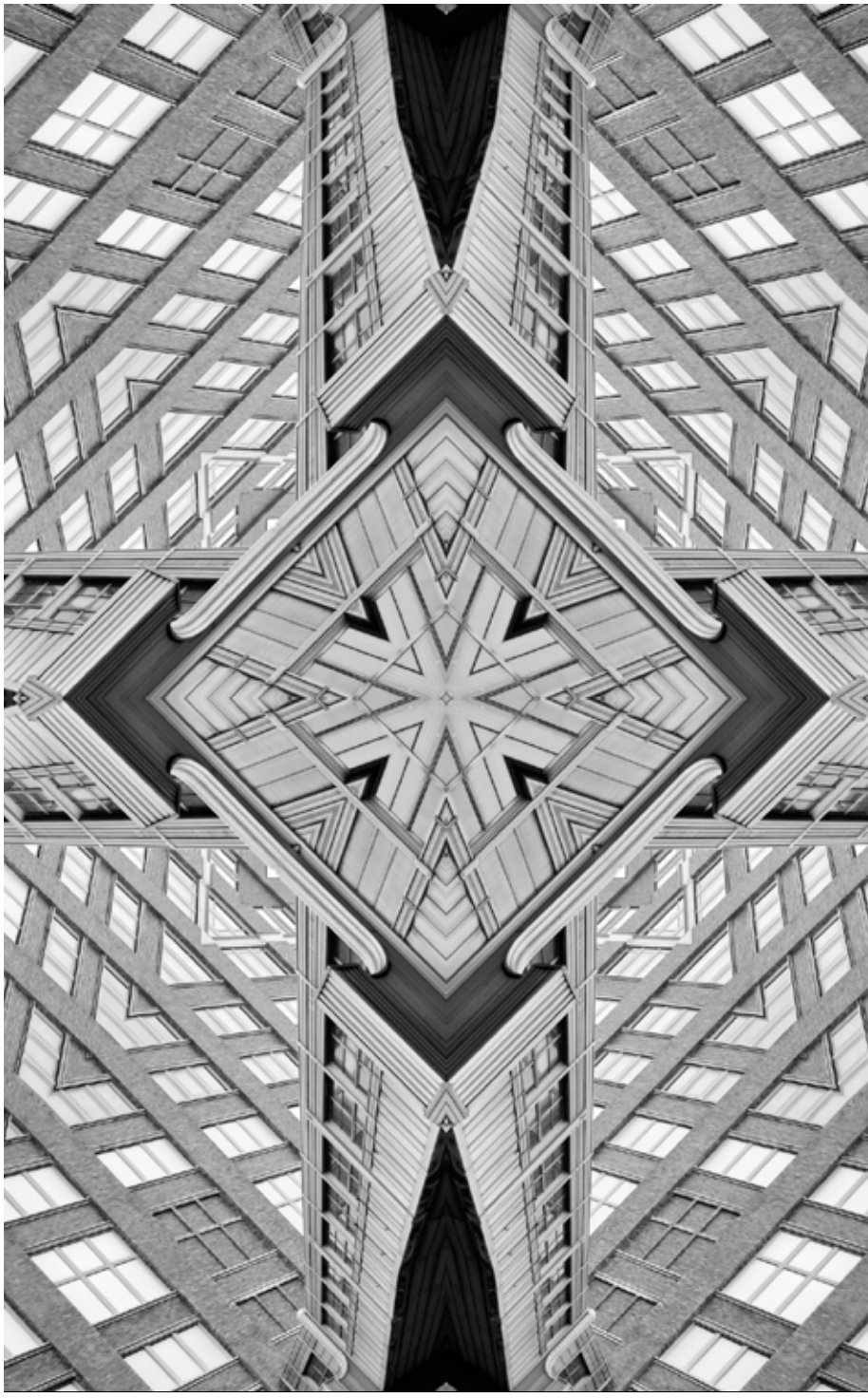
Issue Brief

ISSUE NO. 589
NOVEMBER 2022

OBSERVER RESEARCH FOUNDATION



Ethics of A.I.: Principles, Rules, and the Way Forward

Husanjot Chahal

Abstract

In recent years, different research institutions, government bodies, and private entities across countries have issued principles and guidelines for the ethical use of Artificial Intelligence (AI). There is little consensus, however, over universal ethical principles and how to implement them. What are the similarities and differences in AI ethics discussions across geographies, and what are the existing gaps? Crucially, if the larger goal is the ethical development and deployment of AI, are efforts towards codifying and devising high-level ethical AI principles even a fruitful exercise?

Attribution: Husanjot Chahal, "Ethics of A.I.: Principles, Rules, and the Way Forward," *ORF Issue Brief No. 589*, November 2022, Observer Research Foundation.

The Current Landscape of AI Ethics

Artificial Intelligence (AI) is being deployed in ways that touch people's lives, including in areas of healthcare, financial transactions, and delivery of justice. Advances in AI can have profound impacts across varied societal domains, and in recent years, this realisation has sparked ample debate about the values that should guide its development and use.

States and international organisations have reacted to these societal concerns in various ways. Some have formed ad-hoc committees tasked to deliberate and provide recommendations on the subject. Examples include the United States National Artificial Intelligence Advisory Committee (NAIAC) that dispenses advice to the president and various federal officials; the expert group on AI at the Organisation for Economic Co-operation and Development (OECD); the High-Level Expert Group on AI formed by the European Commission; and the Select Committee on AI appointed by the UK Parliament's House of Lords.[1] These bodies have either drafted or are currently drafting policy documents on the ethical, economic, and social implications of advances in AI.

Similar efforts are underway in the private sector. Companies that are at the forefront of AI development like Google, IBM, Intel, Microsoft, and Sony, have released guidelines for developing ethical AI.[2] Some analysts have propounded that these private entities desire to shape the AI ethics domain in ways that either eschew regulation or else meet their own business priorities.[3] Meanwhile, non-profit organisations and professional associations, such as the Institute of Electrical and Electronics Engineers (IEEE), Internet Society, OpenAI, and the World Economic Forum have also issued declarations and recommendations on AI principles and policies.

The multitude of efforts across such diverse stakeholders reflects the need for guidance in AI development. Not only are the organisations that have produced ethical guidelines on AI diverse—the content of such documents is equally wide ranging. Several empirical studies of AI ethical principles have attempted to examine the various topics under discussion across sectors and countries, and to propose how such principles can be implemented in practice.[4] A review of the findings across these studies can offer insights into the scope and potential for a global agreement on the subject of AI ethics, as well as the disagreements therein.

Points of Convergence

Research shows that most of the available ethical guidelines adopted by states, international organisations, and private companies include a discussion of the following five ethical principles: transparency, justice and fairness, responsibility and accountability, privacy, and non-maleficence.[5] These themes were referenced in at least half of the documents analysed across different studies and could indicate some convergence in global thinking on ethical AI.

a. Transparency. The principle of transparency, or the need to have transparent processes in the development and design of AI algorithms, reflects a commitment to increase interpretability, explainability, or other acts of disclosure. It is one of the most prevalent principles in current literature on AI.[6]

b. Justice and fairness. This principle is expressed mainly in terms of fairness and mitigation of unwanted bias, as a caution to the global community that AI may increase inequality and reinforce societal biases if they are not addressed adequately.[7]

c. Responsibility and accountability. There are widespread references to "responsible AI," although the concept of 'responsibility' is rarely defined. Recommendations centred on responsibility include clarifying legal liability, focusing on underlying processes that may cause potential harm, or whistleblowing in case of potential harm.[8] Responsibility seems to be intertwined with the principles of transparency and justice such that promoting both these themes can increase responsibility and accountability by entities that develop and deploy AI.

d. Privacy. While often undefined, privacy is viewed both as a value to uphold and as a right to be protected in ethical AI, and gets presented commonly in relation to data protection and data security.[9]

e. Non-maleficence. The mention of non-maleficence (encompassing calls for safety and security) exceeded that of beneficence, indicating the precedence of moral obligation to preventing harm over the promotion of good.[10] This could be due to a negativity bias in characterisation of ethical values concentrating more on negative issues and events rather than positive ones.[11] For instance, existing guidelines do not generally discuss how ethical principles could be promoted through responsible innovation in AI.

Points of Divergence

There are substantive divergences across various ethical AI guidelines as analysed by scholars. Most of them relate to the following three main factors:

Interpretation

There are significant differences in how the same principles are interpreted across various guideline documents and the requirements considered important for their realisation. For instance, the need for more datasets to "unbias" AI—to ensure that AI models are trained on representative data in order to avoid flawed or biased conclusions and recommendations—appears to be in conflict with the need to give individuals greater control over their data and ensure privacy. Some guidelines emphasise the need to balance risks and benefits in AI development while others talk of avoiding harm at all costs.[12]

Attribution

There are also divergences in attribution—interpreting which domain, actor, or issue these ethical principles pertain to. For instance, does the European guideline on privacy (encompassing protection of individual's data from both state and commercial entities) also apply to China where privacy guidelines target only private companies, and citizens are accustomed to living in a protected society with high trust in their government?[13] Different perspectives, interpretations, and priorities in ethical AI are of course to be expected given that these documents are developed by a broad range of countries, international organisations, and companies. That said, such divergences could undermine attempts to develop a global ethical AI agenda because varied perspectives, for example risk-benefit evaluations, will lead to different results based on whose well-being they are developed for or the actors involved in developing them.[14]

Implementation

Finally, there are differing opinions on how ethical AI principles should be implemented—through government organisations, inter-governmental organisations, industry leaders, individual users or developers, or by harmonising AI agendas across the board. If harmonisation is a goal, then how does one account for moral pluralism and cultural diversity across countries, considering that AI is a general-purpose technology operating in varied contexts and cultures?

Persistent Gaps

Discussions on the ethical development and use of AI are ongoing, and as such, there are gaps that remain unaddressed. For example, themes of sustainability and solidarity are sparsely referenced across documents.[15] Sustainability appears more commonly in public sector documents than in those drafted by private or non-governmental organisations (NGOs).[16] AI deployment today requires massive computational resources, and thus high energy consumption, and this need will only expand with time. This makes the broader underrepresentation of sustainability-related principles particularly concerning, and calls into question the possibility of harnessing the benefits of AI for the entire biosphere.[17]

Solidarity—a concept mostly referenced in relation to the consequences of AI for the labour market—is also absent in most discussions. There are very few guidelines that pay attention to promoting solidarity by exploring the use of AI expertise for redistributing the augmentation of prosperity for all, and solving socio-economic challenges such as job losses, inequality, and unfair sharing of burdens. Sharing prosperity could mean, for example, compensating humans whose actions provide data for training AI models.[18]

Integrity—meaning being explicit about best practices and disclosure of errors—is another theme that is missing across guideline documents.[19] Current documents place crucial focus on propagating the values of accountability and responsibility, but hardly any emphasise the duty of all stakeholders to develop and deploy AI with integrity. Similarly, the discussion of lack of diversity within the AI community is mostly absent, which is problematic because such dearth of diverse thought could result in flawed AI systems that perpetuate gender and racial biases.[20]

Several initiatives, particularly those offered by industry, are generally criticised as mere virtue-signalling designed to debate on abstract problems and delay regulation.[21] In relation to this, it has been observed that many guidelines, especially those produced by the private sector, indicate that technical solutions exist for several of the identified issues, such as privacy and non-maleficence. However, very few guidelines have offered, or at least acknowledged, technical explanations at all; and when they do, they are sparse.[22] While one cannot expect guidelines to be exhaustive about all problems AI could cause, issues

Persistent Gaps

pertaining to political abuse of AI systems—generating election fraud, fake news, and propaganda, which are widely acknowledged as critical problems of today—are also an oversight.

Furthermore, shifting the focus from principle-development to implementation is an important next step. However, existing discussions lack clarity on which ethical principles should be emphasised and how existing conflicts in interpretation can be resolved. Moreover, there is a need to determine how conflicts between ethical principles can be resolved and who should enforce oversight and ensure that researchers and institutions comply with ensuing guidelines.

> "Themes of sustainability and solidarity are sparsely referenced in documents on AI ethics released so far by states, NGOs, and private entities."

Factors for the Convergence, Divergence, and Gaps

The field of AI ethics is expanding. Convergences across the five ethical principles is understandable as it could be a testimony to the significance of those principles; divergences likely reflect the diversity in viewpoints; and gaps could result because much of the work in this domain is still in progress. Having said that, it is crucial to consider other factors possibly influencing these results.

A significant question pertains to equality of participation in the ongoing global discussion on AI ethics. Some scholars have indicated that the current AI ethics discourse is mostly dominated by countries in the Global North.[23] Indeed, of the 506 AI-related documents listed in Council of Europe's data visualisation of AI initiatives (as of October 2022), only 10 percent come from countries outside Europe and North America.[24] Moreover, research indicates that there is a dearth of reference to key terms associated with gender within AI ethics documents and the ratio of female-to-male authors across these documents is a low 31 percent.[25] Therefore, like other parts of AI research, the discourse on AI ethics is also primarily shaped by men. The absence of an inclusive AI ethics landscape means that mainstream discussions are reinforcing certain viewpoints while possibly neglecting other risks and ethical considerations of importance to women and countries beyond Europe and North America.

Consensus or dissensus among AI ethics documents could also result due to the provenance of literature. Different types of organisations—public, private, and NGOs—have differing priorities, audiences, motivations, and scope of responsibility. The public sector is known to emphasise questions related to unemployment and economic growth, while the private sector focuses more on ethical issues with technical fixes (such as transparency and algorithmic bias); for their part, NGOs address a broader range of topics such as accountability and misinformation.[26] In comparison to the private sector, NGOs and public sector entities are reportedly more similar to each other in their approach to AI ethics—they have more participatory processes in creation of guidelines, greater engagement with issues of regulation and law, and more depth and ethical breadth.[27] Consequently, depending on the corpus of documents and types of organisations at hand, an assessment of AI ethics could indicate meaningful variations or similarities in the choice of topics.

In AI ethics, what forms "AI for good" is under negotiation through dialogues among people or organisations impacted by AI development and other intergovernmental initiatives. If calls for more technology access and multi-stakeholder participation are followed, the field is likely to become even more diverse. Narrower versions of the existing themes are likely to emerge with respect to particular geographies and stakeholder groups.[28] This strengthens the case for putting more effort into clarifying the variations that exist within themes and also undertaking measures to resolve differences in interpretation or attribution where possible. If the goal is to have a better articulated ethical AI landscape, the current discourse should be enriched through evaluation of critical but underrepresented principles, such as sustainability and solidarity underlining social and ecological costs of AI.

Beyond a principled approach to AI ethics

While 'principlism' has been the underlying framework to influence the development of safe and beneficial AI, many have questioned its effectiveness. Some critics have pointed out that the field of AI ethics has produced largely vague and high-level principles and value statements. A 2018 study by McNamara et al. reviewed the idea that ethical guidelines serve as a basis for ethical decisions made by developers.[29] The study found that the effectiveness of guidelines is almost negligible since it does not change the behaviour of students or technology professionals.

Relatedly, scholars have indicated that there are other reasons to be concerned about the future impact of AI ethical guidelines.[30] Certain characteristics of AI development indicate that any principled efforts at ethics might not have significant impact on AI's governance and design.

First, the fundamental aims of AI developers, users, and affected parties do not align, and a unified regulatory framework does not exist yet in the field that establishes clear fiduciary duties towards data subjects and users. This means that users cannot trust that developers will act in their best interests when implementing ethical principles in practice. Reputational risks may compel companies, and personal moral conviction may press AI developers towards good behaviour. However, any righteous actions that place public interests before the company and that do not align with company incentive structures are unlikely.[31]

Second, the situation gets further complicated given that AI development lacks a homogenous professional culture, history, moral obligations, and professional standards of what it means to be a "good" AI developer. AI ethics initiatives try to address this gap by offering broadly acceptable guidelines for AI development across radically different contexts of use.[32] But this results in principles or values that are abstract and based on vague concepts that are not specific enough to guide action and are left to developers to interpret as they see fit.

Third, outside of academic contexts, any principled approach to AI ethics does not have proven methods to transform principles into practice. For instance, the field of medicine has numerous professional societies, accreditation and licensing boards, ethics review bodies, codes of conduct, peer self-governance, and other mechanisms reinforced by strong institutions that ensure ethical conduct on a daily basis.[33] AI development lacks comparable structures to translate guidelines into practice to ensure that this technology, developed behind closed doors, is value-conscious.

Finally, a key weakness for AI is the relative lack of professional and legal accountability mechanisms to redress misbehaviour and ensure that standards are upheld. Research indicates that the existence of mere codes of ethics is not sufficient, and they are often viewed as "checklists" that get pursued in letter rather than spirit.[34] Broader guidelines and self-regulatory efforts alone cannot prevent AI development from failures or misuse, and existing norms and requirements will not be able to set matters right. What makes matters more complicated is that setting up strong accountability mechanisms in AI appears unlikely in the future given that AI is not a unified profession operating in a single sector with a long history of harmonised aims. All of this questions the need for high-level principles as a tool to effect change.

Conclusion

A plethora of national, international, and commercial AI guidelines in recent years have paved the way for some progress on the development of a principles-led approach to AI. However, one should not celebrate limited consensus on high-level guidelines that conceal deep normative and political disagreements.[35] Instead, it is time to move forward in defining clear long-term pathways, setting explicit professional standards tailored towards specific applications, and building accountability structures that are not only country-specific but also sector- and organisation-specific. Mechanisms should also be set up to license developers of applications with elevated risks, such as facial recognition tools or other systems trained on biometric data.

It will also be interesting to see any future AI principles-based discussions geared toward particular applications of AI, like autonomous vehicles, credit scoring services, recruitment procedure software, or other high-risk AI. There have been instances where ethically motivated efforts have been undertaken to improve AI systems, and most of them have been in specific fields where technical solutions exist for particular problems. For example, many privacy-preserving techniques, like homomorphic encryption or federated learning, or other methods using differential or stochastic privacy, have been developed for the use of data and learning algorithms.[36] A deeper assessment of these context-specific cases to underline guidelines for AI principles could be a way forward.

Admittedly, principles are difficult to translate into practice. However, they still play a crucial role in building awareness and acting as catalysts for building beneficence and a culture of responsibility among AI developers. Internalised norms and values have a role in influencing extrinsic measures, and how individual developers conceptualise, communicate, and enforce extrinsic measures will be crucial in facilitating their implementation. Principles alone cannot govern AI, but nor can rules and requirements.[37] An effective AI governance strategy will require both—principles encouraging cultural change in the AI community, and explicit rules and regulations buttressing them.

(This brief was first published in Digital Debates 2022, *ORF's annual journal on technology and society.)*

Husanjot Chahal *is a Research Analyst at Georgetown University's Center for Security and Emerging Technology (CSET).*

Endnotes

1 U.S. National Artificial Intelligence Initiative, *The National AI Advisory Committee (NAIAC)*, (Washington, D.C: 2022), https://www.ai.gov/naiac/; Organisation for Economic Co-operation and Development, *OECD creates expert group to foster trust in artificial intelligence*, (2018), https://www.oecd.org/innovation/oecd-creates-expert-group-to-foster-trust-in-artificial-intelligence.htm; European Commission, *High-level expert group on artificial intelligence*, https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai; United Kingdom Parliament Select Committee on Artificial Intelligence, (London: 2017), https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10001.htm.

2 Google AI, "Artificial Intelligence at Google: Our Principles," Google, https://ai.google/principles/; IBM Think Blog, "Transparency and Trust in the Cognitive Era," 2017, https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/; Intel, "Artificial Intelligence: The Public Policy Opportunity," 2017, https://community.intel.com/legacyfs/online/files/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf; Microsoft, "Responsible AI," https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6; Sony, "Sony Group's Initiatives for Responsible AI," https://www.sony.com/en/SonyInfo/sony_ai/responsible_ai.html.

3 Daniel Greene et al., "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning" (paper presented at the Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019), https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/849782a6-06bf-4ce8-9144-a93de4455d1c/content.

4 Fjeld et al. compared 36 documents side by side to identify trends that suggest the earliest emergence of sectoral norms. Zeng et al. collected 27 proposals of AI principles and introduced Linking Artificial Intelligence Principles (LAIP), a platform to link and analyze them. Jobin et al. conducted a scoping review of the existing corpus and analyzed 84 documents of AI ethical guidelines in their paper; Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI," Berkman Klein Center for Internet & Society, 2020, https://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf?sequence=1&isAllowed=y; Yi Zeng et al., "Linking Artificial Intelligence Principles" (paper presented in the Proceedings of the AAAI Workshop on Artificial Intelligence Safety, AAAI-Safe AI, 2019), https://arxiv.org/pdf/1812.04814.pdf; Anna Jobin et al., "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, 389-399 (2019), https://www.nature.com/articles/s42256-019-0088-2.

5 Jobin et al., "The global landscape of AI ethics guidelines"

6 It featured in 73 out of 84 sources analyzed by Jobin et al. and 94 percent of the documents in Fjeld et al.'s dataset; Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI"

Endnotes

7 Jobin et al., "The global landscape of AI ethics guidelines"

8 Official Microsoft Blog, "Responsible bots: 10 guidelines for developers of conversational AI," Microsoft Corporation, 2018, https://www.microsoft.com/en-us/research/publication/responsible-bots/; Christina Demetriades and Tom McLaughlan, "Responsible AI and Robotics: An ethical framework", Accenture, https://www.accenture.com/gb-en/company-responsible-ai-robotics.

9 Australian Government, Commonwealth Scientific and Industrial Research Organisation, *Artificial Intelligence: Australia's Ethics Framework CSIRO Data61 report: Artificial Intelligence: Australia's Ethics Framework*, (Canberra, Australia: 2019),https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjV4Yziiqn6AhWQM1kFHS3RC2kQFnoECAkQAQ&url=https%3A%2F%2Fwww.industry.gov.au%2Fpublications%2Faustralias-artificial-intelligence-ethics-framework&usg=AOvVaw3wz-VNTEhbMSjjq6uW7Acm; Sony, *Sony Group's Initiatives for Responsible AI*; Intel, *Artificial Intelligence: The Public Policy Opportunity*.

10 Harm is generally interpreted across documents as discrimination, violation of privacy, or physical harm.

11 Thilo Hagendorff, "AI ethics and its pitfalls: not living up to its own standards?" *AI and Ethics*, (2022), https://link.springer.com/article/10.1007/s43681-022-00173-5

12 Jobin et al., "The global landscape of AI ethics guidelines"

13 Pascale Fung and Hubert Etienne, "Can China and Europe find common ground on AI ethics?," *World Economic Forum*, 2021, https://www.weforum.org/agenda/2021/11/can-china-and-europe-find-common-ground-on-ai-ethics/.

14 Jobin et al., "The global landscape of AI ethics guidelines"

15 Some of the documents that mention sustainability (protecting the environment, improving biodiversity, minimizing ecological footprint, creating fairer and equal societies, etc.) are by the Future of Life Institute, Green Digital Working Group, and the French Parliamentary mission. Solidarity is referenced in varied contexts (implications of AI for labor market, calls for a strong safety net, etc.) by the Norwegian Data Protection Authority, U.S. National Science and Technology Council, and by academic researchers at various universities.

16 As per Cathy Roche et al. (2021), the term "sustainable" has been referenced 18/31 times in documents by the public sector, 8/35 times by NGOs, and 2/18 times in private sector documents; Cathy Roche, Dave Lewis and P. J. Wall, "Artificial Intelligence Ethics: An inclusive global discourse?" *arXiv* (2021), https://arxiv.org/pdf/2108.09959.pdf.

17 Sergio Genovesi and Julia Maria Mönig, "Acknowledging Sustainability in the Framework of Ethical Certification for AI," *Sustainability*, 14(7), 4157, (2022), https://doi.org/10.3390/su14074157.

Endnotes

18 Miguel Luengo-Oroz, "Solidarity should be a core ethical principle of AI," *Nature Machine Intelligence 494*, 2019, https://www.nature.com/articles/s42256-019-0115-3.

19 Valerie Carey, "AI Integrity: Leadership Lessons from Other Industries," *Towards Data Science*, February 4, 2022, https://towardsdatascience.com/ai-integrity-leadership-lessons-from-other-industries-82e3d6af2e95.

20 Kelsey Snell, "Lack of diversity in AI development causes serious real-life harm for people of color," *NPR*, February 13, 2022, https://www.npr.org/2022/02/13/1080464162/lack-of-diversity-in-ai-development-causes-serious-real-life-harm-for-people-of-.; Maria Klawe, "Why Diversity in AI Is So Important," *Forbes*, July 16, 2020, https://www.forbes.com/sites/mariaklawe/2020/07/16/why-diversity-in-ai-is-so-important/?sh=1c64456c7f2b.

21 Brent Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, Volume 1, 501-507 (2019), https://www.nature.com/articles/s42256-019-0114-4.

22 Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines*, 30, 99-120 (2020), https://link.springer.com/article/10.1007/s11023-020-09517-8.

23 The United States and the United Kingdom cumulatively contributed to 40 percent of all the 84 documents analysed by Jobin et al. study. Cathy Roche et al., "Artificial Intelligence Ethics: An inclusive global discourse?"

24 Council of Europe, *AI initiatives*, https://www.coe.int/en/web/artificial-intelligence/national-initiatives

25 Roche et al., "Artificial Intelligence Ethics: An inclusive global discourse?"; Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines."

26 Daniel Schiff et al., "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection," *TechRxiv* (2021), https://www.techrxiv.org/articles/preprint/AI_Ethics_in_the_Public_Private_and_NGO_Sectors_A_Review_of_a_Global_Document_Collection/14109482/1.

27 Daniel Schiff et al., "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection."

28 Jessica Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI"

29 Andrew McNamara et al., "Does ACM's code of ethics change ethical decision making in software development?" (paper published in Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018), https://dl.acm.org/doi/10.1145/3236024.3264833.

30 Brent Mittelstadt, "Principles alone cannot guarantee ethical AI"

Endnotes

31 Jennifer J. Kish-Gephart et al., "Bad apples, bad cases, and bad barrels: meta-analytic evidence about sources of unethical decisions at work," *Journal of Applied Psychology*, 95, 1–31 (2010), https://www.researchgate.net/publication/41087509_Bad_Apples_Bad_Cases_and_Bad_Barrels_Meta-Analytic_Evidence_About_Sources_of_Unethical_Decisions_at_Work; Daisuke Wakabayashi and Scott Shane, "Google will not renew Pentagon contract that upset employees," *The New York Times*, June 1, 2018, https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html.

32 Brent Daniel Mittelstadt, "The ethics of algorithms: Mapping the debate," *Big Data & Society* (2016), https://journals.sagepub.com/doi/10.1177/2053951716679679.

33 Stephen Toulmin, "How medicine saved the life of ethics," *Perspectives in Biology and Medicine, 25*, 736–750 (1982), https://muse.jhu.edu/article/404227.

34 Livia Iacovino, "Ethical principles and information professionals: theory, practice and education," *Australian Academic & Research Libraries, 33*, 57–74 (2002), https://www.tandfonline.com/doi/abs/10.1080/00048623.2002.10755183; Rosamond Rhodes, "Good and not so good medical ethics," *Journal of Medical Ethics, 41*, 1 (2015), https://jme.bmj.com/content/41/1/71.info.

35 Brent Mittelstadt, "Principles alone cannot guarantee ethical AI"

36 John C, Duchi et al., "Privacy Aware Learning," *arXiv* (2013), https://arxiv.org/pdf/1210.2085.pdf; Benjamin Baron and Mirco Musolesi, "Interpretable Machine Learning for Privacy-Preserving Pervasive Systems," *arXiv* (2020), https://arxiv.org/pdf/1710.08464.pdf.

37 Elizabeth Seger, "In Defence of Principlism in AI Ethics and Governance," *Philosophy & Technology*, 45 (2022), https://link.springer.com/article/10.1007/s13347-022-00538-y.

Images used in this paper are from Getty Images/Busà Photography.

ORF OBSERVER RESEARCH FOUNDATION

Ideas . Forums . Leadership . Impact

20, Rouse Avenue Institutional Area,
New Delhi - 110 002, INDIA
Ph. : +91-11-35332000. **Fax :** +91-11-35332005
E-mail: contactus@orfonline.org
Website: www.orfonline.org