AI and Ethics (2021) 1:131–138
https://doi.org/10.1007/s43681-021-00052-5

OPINION PAPER

Towards intellectual freedom in an AI Ethics Global Community

Christoph Ebell[1,2] · Ricardo Baeza-Yates[3] · Richard Benjamins[4] · Hengjin Cai[5] · Mark Coeckelbergh[6] · Tania Duarte[7] · Merve Hickok[8] · Aurelie Jacquet[9] · Angela Kim[10] · Joris Krijger[11] · John MacIntyre[12] · Piyush Madhamshettiwar[13] · Lauren Maffeo[14] · Jeanna Matthews[15] · Larry Medsker[16] · Peter Smith[12] · Savannah Thais[17]

Received: 15 March 2021 / Accepted: 20 March 2021 / Published online: 13 April 2021

Abstract
The recent incidents involving Dr. Timnit Gebru, Dr. Margaret Mitchell, and Google have triggered an important discussion emblematic of issues arising from the practice of AI Ethics research. We offer this paper and its bibliography as a resource to the global community of AI Ethics Researchers who argue for the protection and freedom of this research community. Corporate, as well as academic research settings, involve responsibility, duties, dissent, and conflicts of interest. This article is meant to provide a reference point at the beginning of this decade regarding matters of consensus and disagreement on how to enact AI Ethics for the good of our institutions, society, and individuals. We have herein identified issues that arise at the intersection of information technology, socially encoded behaviors, and biases, and individual researchers' work and responsibilities. We revisit some of the most pressing problems with AI decision-making and examine the difficult relationships between corporate interests and the early years of AI Ethics research. We propose several possible actions we can take collectively to support researchers throughout the field of AI Ethics, especially those from marginalized groups who may experience even more barriers in speaking out and having their research amplified. We promote the global community of AI Ethics researchers and the evolution of standards accepted in our profession guiding a technological future that makes life better for all.

1 Introduction

We are living in the era of Artificial Intelligence (AI) where almost everyone in society is interacting with AI embedded services or products, whether knowingly or unknowingly. In addition, AI poses unique challenges due to its scalability and exponential growth in the number of applications in society. We view AI Ethics research as being in an early stage, characterized by awareness of issues such as trustworthiness, transparency, accountability, diversity, and non-discrimintaion along with initial ideas on practical ways to address ethical problems. While we are developing rapidly as a field and in this context, it is important to note that AI Ethics is at a crossroads. Like all fields of knowledge, it is

✉ Christoph Ebell
christoph.ebell@verticai.org

Tania Duarte
http://www.weandai.org

Merve Hickok
http://www.Alethicist.org

[1] HumanIn Association, Paris, France

[2] Verticai Consulting, Geneva, Switzerland

[3] Institute for Experiential AI, Northeastern University, Boston, Massachusetts, USA

[4] Observatorio del impacto social y ético de la inteligencia artificial, Barcelona, Spain

[5] School of Computer Science, Wuhan University, Wuhan, People's Republic of China

[6] Universität Wien, Wien, Austria

[7] We and AI, London, UK

[8] AIethicist.org, Ann Arbor, MI, USA

[9] Lawyer, Sydney, Australia

[10] Women in AI, Sydney, Australia

[11] Erasmus Universiteit Rotterdam, Rotterdam, The Netherlands

[12] Sunderland University, Sunderland, UK

[13] Department of Transport, Victoria, Australia

[14] Washington, DC, USA

[15] Clarkson University, Potsdam, USA

[16] George Washington University, Washington, DC, USA

[17] Princeton University, Princeton, USA

subject to power relations, political interests, and business objectives such as quarterly earnings goals. We are making progress on the ethical basics, but when it comes to application production and implementation, we face known issues of who writes the narrative and defines the discursive frameworks. Ethics is increasingly discussed at AI conferences and in journals, and many researchers in academics and technology companies are creating frameworks and technical processes for detecting AI ethical issues in data collection, model building, and implementation processes, but we have substantial work to do as a field.

The firings of Drs. Timnit Gebru and Margaret Mitchell from Google [1–3] bring into sharp focus many of the challenges we face as a field and make clear the need to establish an AI Ethics culture and safe processes in the workplace for AI Ethics researchers [4]. We note a number of relevant aspects of this highly publicized case:

- Internal communications revealed that at least three authors of an AI Ethics-related paper, including Gebru and Mitchell were told to refrain from casting Google's technologies in a negative light [2, 3].
- The research results in their paper focused on large language models such as the Bidirectional Encoder Representations from Transformers (BERT) [5], which affects billions of people as part of Google's search engine.
- Multinational companies such as Google and others can have massive impacts across the globe and leverage their market position and resources across multiple jurisdictions.
- Bias and discrimination are recognized, serious problems related to the use of AI; systemic racism across many organizations creates factors that are involved in undermining diverse, and particularly black, voices. A much more systematic and thorough analysis is needed to understand them and how they interact and create negative impacts. Such factors have been reported in news sources as being lack of promotion opportunities, inequitable remuneration, lack of sponsorship, incidences of harassment, inequitable HR practices, and more rigid disciplining, some of which are reportedly extended to their internal allies. It is necessary to understand to what extent these factors are present despite the diversity, equity and inclusion programmes undertaken by organisations, and why such programmes fail to prevent environments being seen as hostile to black and other underrepresented employees.
- Big Tech corporations are employing a significant part of the best researchers around the world (e.g., Google, Apple, Facebook, Amazon, and Microsoft (GAFAM) hired about 100,000 employees coming from top US universities [6]), and therefore they have acquired a certain responsibility for independent science, a role which traditionally is played by universities and other academic institutions.
- Certain minority groups are significantly underrepresented in the workforce at GAFAM companies (and generally in roles relating to AI), risking a lack of regard for the value of different perspectives and lived experiences within research projects, and resulting in inequitable work cultures and practices which place minority researchers at greater risk of suppression.

In 2013, the European Environment Agency (EEA) explained in a report [7] that many of its case studies revealed "the problems faced by early warning scientists who have been harassed for their pioneering work, including bans on speaking out or publishing, loss of funding, legal or other threats, and demotion" and concluded that one obvious solution is "that scientists in these situations should receive better protection either via an extension of 'whistle blowing' and discrimination laws, or by independent acknowledgement of the value of their work". This statement rings true today also with regards to AI Ethics research. In 2020, Gebru stated in an interview "I don't think the lesson is that there should be no AI ethics research in tech companies, but I think the lesson is that a) there needs to be a lot more independent research and b) there needs to be oversight of tech companies" [2].

Now is the time to provide better protection to AI ethics researchers and scientists as a first concrete step in operationalizing recent AI principles, and that includes improving whistleblowing protections. Whistleblowers are now a recognized essential mechanism for uncovering unethical behavior such as fraud [8], and while recent improvements have been made to whistleblowing laws, we need to adopt unified whistleblowing laws to simplify and facilitate the existing complex legal process we face [7], including an equal level of protection and remedies against retaliation. Collecting data on ethical attributes of companies could identify correlations with stock values to test the importance of ethical reputations of corporations. A large part of AI research is done by industry [6], so AI ethics research must also be conducted there, not just in academic research institutions.

2 A (very) short history of AI ethics

During the past decade, much attention has been given to ethics of AI, in academia (e.g., Benjamin [8]; Bodington [9]; Bostrom [10]; Buolamwini [11]; Coeckelbergh [12]; Mittelstadt [13]; Noble [4]; Wachter et al. [14]), in public policy (e.g., European Commission, UK House of Lords, U.S., Korea, China), and in the corporate world, where large corporations such as Google and Microsoft have taken

initiatives around AI ethics that focus on self-regulation. This has delivered more public attention to ethics of AI and has produced many frameworks with principles for ethical AI. However, some of these initiatives in the public and private sector have been called "ethics washing" and even "ethics shopping" (Wagner [15]): ethical frameworks and initiatives are then used to avoid regulation and emphasize the role of the private sector. Nevertheless, companies and governments have hired ethicists who often do well-intended work aimed at more ethical and responsible AI within their organizations.

In theory, these ethicists can take at least two roles:

1. One is constructive and aims to implement ethics into the technological development or
2. Another one is that of whistleblowers.

In 2018, more CEOs were dismissed for ethical lapses than for financial performance or board struggles [16] and thus heightening concern in corporate leadership for discussions of AI Ethics. The discussion on the social responsibilities of businesses and tech companies has only intensified with the increased use of AI technology. As many guidelines for ethical development and deployment of AI systems take shape and are formalized into technological solutions (e.g. the IBM Explainability and Fairness 360 tools), questions arise about how industry can ensure that AI ethical frameworks can be translated into actual alterations of innovation policies and strategic decision making. Many business leaders acknowledged that AI will substantially transform their industries and agreed on ethics as a dominant concern regarding this transformation [17–20, 23]. Since then, the attention for ethical concerns has only increased and industry leaders are either considering or adopting policies about AI ethics, although the way frameworks influence business practices is often opaque.

In October 2020, 31 scientists wrote a scathing letter to the science journal Nature [21]. The journal had published a study earlier that year chronicling Google Health's account of an AI system that searched medical images to diagnose breast cancer. In their letter, the scientists argued that this study was simply a commercial for proprietary tech, rather than being a rigorous study. "When we saw that paper from Google, we realized that it was yet another example of a very high-profile journal publishing a very exciting study that has nothing to do with science," wrote Benjamin Haibe-Kains, the lead author of the response. "It's not about this study in particular—it's a trend we've been witnessing for multiple years now that has started to really bother us." A primary problem with the study was its lack of reproducibility. In their article, the researchers shared so few details about how the AI system achieved its outcome that they could not try to replicate its results. Reproducibility weeds out studies of lower caliber and is a key part of publication science.

To achieve reproducibility in AI, researchers must provide transparency into specific aspects of how their models were trained. These specificities will vary by use case, so given the nuances of each model, researchers need to share how the models were trained and any barriers in the development process. This pragmatic part of AI Ethics, seeking to apply the findings and concerns to real-world phenomenology, is the focus of our current article.

3 Some key problems

AI Ethics professionals are becoming actively engaged in discussing organizational responsibility [22–25, 25–36, 38, 39, 44], and we encourage even more widespread attention to issues such as the following:

a. A fundamental conflict of interest exists in corporate in-house and corporate-funded AI Ethics research. We can as a community raise and make visible how AI ethics research is seen in the industry, academia, and the broader civil society. Several recent works have explored the tension between the goals of private corporations and unbiased AI ethics research. Indeed, if an ethical AI analysis uncovers potential issues with the sponsoring company's own product, the company has little incentive to acknowledge the issue and therefore provides little oversight on its resolution. This can be attributed in part to the lack of a national or global framework for ethical AI and lack of formalized regulation. To expect corporations to conform to ethical standards is naïve when those standards are not agreed upon or societally binding. Thus, we must take a multi-faceted approach not hinging entirely on corporate self-governance. Independent audits of AI systems have emerged as a critical component of measuring issues such as systematic biases, and some work has shown that these audits result in quantifiable reductions in bias [25, 26]. As a professional community, we should encourage corporations to allow their AI ethics researchers to work with independent researchers to undertake these audits with the expectations of publishing the results. Similarly, we should advocate for more concrete definitions of algorithmic harm and appropriate governmental policy, as this will provide additional external pressure on corporations to prioritize AI ethics work and bring additional public attention to these issues. AI must be seen societally and legally as a profession with fiduciary duties to the public, much like how the field of medicine has a strong history of professional ethical norms [7].

b. Problems of systemic racism and other culturally deeply encoded discriminatory mechanisms multiply the severity of such conflicts of interest insignificant, perhaps non-linear and even counter-intuitive ways [22, 24, 36, 38, 39, 44]. AI Ethics research becomes a site where both problems intersect. An AI Ethics researcher carries a responsibility to surface such pernicious dynamics and is therefore particularly vulnerable to being attacked from multiple angles with exactly the mechanisms they are tasked to uncover and try to remedy. These efforts will significantly aid in ensuring a robust AI ethics research community, but we cannot ignore the intersection of AI ethics with issues of systemic racism and other culturally encoded discriminatory mechanisms. It is well known that racial minorities, women, LGBTQ + individuals, and other groups are underrepresented in AI research and other key technical roles [24]. This means that often AI ethics researchers raising concerns about these populations are raising them to people who are not directly affected by the issue and may, in fact, have other personal or corporate stakes in the project. The nature of AI ethics work necessitates being able to analyze and critique systems of power, and this act often puts researchers in a highly vulnerable position if their work is not properly contextualized or responded to. This is true of AI ethics researchers within industry and those in academia, who can still face personal and professional consequences for speaking out.
c. The urgency of ensuring that AI Ethics researchers can do their jobs for the benefit of everyone means they need an extra layer of protection, independent of where they work. A serious reflection on normal corporate research environments would be very important. The fact that corporations check scientific publications for, from any perspective, sensitive material is a general practice. Many corporate researchers are used to that– edits happen and sometimes publications are blocked. So, scrutinizing a research publication by a corporation is not the main point and not necessarily always a problem per se.

4 What is at stake?

Potentially enormous negative and lasting impacts on a variety of stakeholders can be described in the following three key areas:

4.1 Danger of dividing the AI Ethics research community

The 2020 NeurIPS Conference (although itself sponsored mainly by Big Tech) held a “Resistance to AI Workshop” with Dr. Gebru and other vocal researchers concluded that, "AI research has been concentrating power in the hands of governments and companies and away from marginalized communities” [25]. In a recent public departure from collaboration, Access Now has resigned from Partnership on AI [39], stating that they “did not find that Partnership on AI to Benefit People and Society (PAI) [34] influenced or changed the attitude of member companies or encouraged them to respond to or consult with civil society on a systematic basis.”. PAI was established in 2016 by Apple, Amazon, DeepMind and Google, Facebook, IBM, and Microsoft to “study and formulate best practices on AI technologies, to advance the public’s understanding of AI, and to serve as an open platform for discussion and engagement about AI and its influences on people and society.” Access Now’s departure is not the first criticism. In 2018, some Media Lab employees raised concerns that “PAI’s association with ACLU, MIT and other academic/non-profit institutions practically ends up serving a legitimating function. Neither ACLU, MIT, nor any non-profit has any power in PAI”. The threat, therefore, is a split between organizations representing the interests of civil society and its diverse community and the technological actors themselves. Instead of informing each other’s research and reference systems, there is a risk of creating echo chambers on both sides.

4.2 Independence and direction of research

Requirements attached to the funding of both the in-house and external research, as well as the control of dissemination research, has the potential to impact not only the independence of research but also the direction of research and innovation. Corporate-funded research usually comes with the requirement that projects be aligned with a company’s business interests and can be protected by non-disclosure agreements. GAFAM and defense companies employ a significant part of the best researchers around the world and fund a considerable part of top AI researchers [3]. As funders, they can decide what should and should not be researched. Corporate funding does not imply that researchers could not act with integrity or ethics; however, the gatekeeping power of corporations in deciding the research questions can cut through the promise of the “do good” motto in Silicon Valley and the fundamental goal of AI ethics to contribute towards a fairer, more equitable society that respects human rights. Margaret Mitchell commented that “If we are researching the appropriate thing given our expertise, and we are not permitted to publish that on grounds that are not in line with high-quality peer review, then we’re getting into a serious problem of censorship” [18]. Gatekeeping can allow these funders to “brush aside the moral complexities” of these emerging technologies. Even the many published AI ethics principles and guidelines rarely question the business culture, revenue models, and incentive mechanisms that continuously push questionable or harmful AI products into the markets [19,

20, 23]. Nor do they put the spotlight on the public–private partnerships and industry-funded research in the field of AI.

4.3 Diversity and representation

Perceived (and often real) negative or even hostile attitudes towards research in AI Ethics stem from people from diverse backgrounds. As another example, many large corporations have developed AI applications which can be used to support disabled people. Day-to-day applications include speech technology to operate the computer, on smart phones and to answer questions, play music and help with general organizational activities. Many of these applications offer great support and value to disabled individuals; however, they also can fail to live up to their promise and be frustrating [45]. Such corporations have an obligation to involve the disabled community in the co-design of AI products – and to reflect on the ethics of doing so. Of course, this is not easy to achieve. Any community is ultimately diverse, complex and difficult to define. However, this does not mean that corporations and technology providers/designers should not attempt to involve as many disabled individuals as possible, covering as many disabilities as possible, in the design of their products and the entire value chain, including AI Ethics research done by researchers with diverse experiences. Inclusion and exclusion are at stake when it comes to who is "allowed" to speak in AI Ethics research. At the same time, inclusion and exclusion are research categories of AI Ethics. The result is a potentially self-replicating and -scaling circular research practice, undermining the very mechanisms AI ethics research sets out to uncover and propose remedies for. There clearly are political implications of this: legacies of inclusion in and exclusion from political and economic power are now at the cusp of being "baked into" technological means of decision support or even autonomous decision-making with a realistic possibility of remedy questionable, once such systems are deployed and widely used in fields such as human resources, policing, the judiciary, education, healthcare and others fundamental to the functioning of society and the rights of its citizens.[1]

5 The way forward

We end this paper with an initial set of concrete, actionable ideas that we urge our global professional community to research and discuss. The future of AI research and development will have huge positive and negative impacts on global citizens and societies, and we must start now to mitigate the dangers. AI Ethics research is fundamental to this.

Given the potential ethical challenges of AI, relevant research communities should be supported for research on the ethical and societal problems raised by AI in a way that attends to problems regarding exclusion and inclusion, creates bridges between academia and corporate initiatives, and implements ethics in the relevant engineering and computer science curricula. An active and aware AI Ethics research community can work to raise the standard of how AI Ethics research and its workers are seen and treated. We propose a strategic research agenda around topics such as:

- AI Ethics research and Societal Problems (exclusion and inclusion).
- Corporate Risk Governance and AI Ethics Research.
- AI Ethics Research in Engineering Curricula.
- AI Ethics Research and De-risking Big Data for AI.
- The Role of AI Ethics Research in Publicly Funded Research and Industry Collaboration.
- How-to Guidelines for Companies incorporating AI Ethics Research.

In particular, we propose the following action points:

a. Make sure all people are treated with respect, researchers or not, inside institutions. That implies to listen to their points of view and be treated ethically, giving a reasonable explanation for decisions that affect the person, within the bounds and limitations of relevant employment law protections.
b. Organize ethical governance, for example by implementing Ethics boards in companies and universities with at least 50% external members, particularly for research projects, user studies, and even products. Organizations should have in place an appropriate governance framework that allows them to address research findings internally or with external peers in a safe and confidential manner.
c. AI ethics research could be mainstreamed by funding agencies (like the NSF in the U.S. or ERC in the EU, MEXT in Japan, NSFC in China, and so on) for example, through requiring a percentage of funding for AI research going to AI Ethics research for a given project.
d. Advocate for companies and in particular large organizations, to make their Ethical AI guidelines and best practices publicly available, and to provide information on compliance, for example in their annual reports.
e. The AI Community could make it the norm that AI Ethics Research comes with a statement of COI and transparency about who funds and who green-lights

[1] This short section was written by one of the authors using speech technology. The technology worked quite well, although there was a need to correct some of the words. Nonetheless, the author could not work, or write, without access to such technology. Hence this is not a criticism, merely a plea to involve disabled people in the co-design of products in the future.

publications and by which criteria. Papers lacking such a statement would increasingly be implicitly acknowledging lack of due diligence.

f. Implementing an independent whistle-blower mechanism in corporate environments where AI Ethics and AI basic research is performed. (For a discussion of whistleblowing in bioethics see [31].)
g. Providing space in academic journals for the "Ethics of AI Ethics Research" and perhaps publishing options for designated "investigative pieces" or "witness pieces" that diverge from the regular academic format and yet can be fact-checked.
h. Actively giving a platform to under-represented, discriminated-against, or censored members of our community, knowing that they take a bigger risk in speaking out, and consequently making an extra effort to support them.
i. Mentorship for AI Ethics Researchers working in challenging settings.
j. Establishing a central mechanism that requires researchers and/or funders to submit the details of research funding to a central (or decentralized blockchain) database when any paper is submitted. We need transparency in corporate, governmental, and academic research funding for evaluating impacts and reach. Published papers will help disseminate needed information, but a broader picture is obtained by the allocation of resources by our institutions and corporations.
k. Explore technological solutions that implement ethical AI. For example, a standard of secure, immutable, and verifiable records of when and for what certain algorithms are used in corporate or public applications (such as HR, healthcare, judiciary, social media) could drastically increase trustworthiness of such records and restore independent verifiability of certain ethics claims. As such this could also serve as a strong incentive for corporations to design their systems with this in mind. Note, however, that DLT is not an easy "tech fix" for a set of problems that are at their roots behavioral and societal ones. DLT comes with its own set of ethical issues and care must be taken to not import these and make things worse rather than better. Here, too, breaking down boundaries between disciplines is the only realistic way forward.

6 Summary and conclusions

Compelled by recent events, we propose a renewed effort by the global community of AI Ethics researchers to promote the positive uses and mitigate the harmful effects of AI research and development. We need clear and actionable standards in our profession to guide a technological future providing better lives for individuals and society. Corporate as well as academic research settings require responsibility, duties, dissent, and conflicts of interest. The article is meant to provide a reference point at the beginning of this decade regarding matters of consensus and disagreement on how to enact AI Ethics for the good of our institutions, society and individuals. We have identified issues that arise at the intersection of the information technology industry, socially encoded behaviours and biases, and individual researchers' work and responsibilities. The legitimate goals of corporate, governmental, and academic research organizations must allow for the governance of AI technologies as they impact our future.

The role of "big tech" companies, both in their development and use of A.I., and their recruitment, employment, and treatment of A.I. researchers, places a great responsibility on them which they must now acknowledge and meet. It has become clear that social media platforms have used A.I.-based pattern-matching algorithms which were originally developed or consumer profiling to target users with information, and misinformation, which is likely to keep them online longer. In September 2020, Tim Kendall, formerly a senior executive, gave testimony to a Congressional Hearing entitled "Mainstreaming Extremism: Social Media's Role in Radicalizing America." [44]. During his testimony, Kendall asserted that the deliberate use of A.I to "addict" users, keeping them online as long as possible to maximise commercial advantage – and that targeting users with misinformation was part of this process. Kendall passionately described how damaging he felt this strategy was to society, and analogised it to the behaviour of "big tobacco" from the 1950s through to the 1980s (and arguably even the present day) who concealed the dangers of smoking and the addictive nature of nicotine whilst at the same time adding ingredients to their products (sugar, menthol) to increase the addiction they denied [22, 45]. There is even a parallel to be drawn with the automotive industry, which resisted introducing safety measures such as seat belts despite overwhelming evidence of the need for such measures from road traffic accident data [46]. They were, finally, forced to confront these issues when "whistleblowers" such as Ralph Nader went public with the information which the automotive manufacturers were suppressing [47].

Will "big tech" become the "big tobacco" or "big auto" of our time? Will social media and browser mega-corporates stand up and take their responsibilities for how they develop and use A.I., and how they treat the researchers they employ—including those who want to raise questions about the impact of these technologies, in ethical and moral terms? It seems the key question is one of trust. Can we trust "big tech" to do the right thing of their own accord—unlike big tobacco and big auto—or must they be forced, through regulation, oversight, and legislation—to

do so? This question of trust is pivotal to the role "big tech" is already playing in our society, and their use of AI is central. The UK's House of Lords recently held a Select Committee under the leadership of Lord David Puttnam, who has addressed this issue of trust directly in his Committee's report "Digital Technology and the Resurrection of Trust" [48]. Lord Puttnam states that we face an even greater threat to society than the blight of the COVID-19 pandemic, namely: "…a virus that affects all of us in the UK—a pandemic of 'misinformation' and 'disinformation'. If allowed to flourish these counterfeit truths will result in the collapse of public trust, and without trust democracy as we know it will simply decline into irrelevance."

Intellectual freedom is a *sine qua non* for AI ethics research to lay the foundations for a trustworthy, safe, and equitable data economy. There may be perceived short-term corporate or state actor advantages gained through enforcing limitations on intellectual freedom. In the medium and long term, however, the negative externalities of such limitations will disproportionately outweigh any gains. It will sow suspicion and doubt where trust must be built and earned.

Because a thriving, diverse, and intellectually rigorous field of AI ethics is of such fundamental importance to safeguarding society's legitimate democratic interests alongside technological advances, its researchers' intellectual freedom requires special protection. While measured regulatory approaches will be necessary and helpful, it is ultimately a strong, open, and communicative global AI ethics community that can raise the standard of how corporations, universities, governments, and other stakeholders treat one of their most valuable assets in a rapidly changing economy. An economy, to be sure, that is built increasingly on and with the help of artificial intelligence. Therefore, it is incumbent upon all of us, all stakeholders, to play a part in coming together as a community, to demand and protect intellectual freedom in AI ethics research, whatever the setting. Explicitly, corporate responsibility towards protecting intellectual freedom of the researchers funded and employed by them cannot be a "nice to have"—it must be enshrined as a matter of principle. We experience a moment in history, and an opportunity to do the right things, that may never return.

Funding This study is not financially supported.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

1. Hao, K.: We read the paper that forced Timnit Gebru out of Google. Here's what it says. MIT Technology Review. (2020). https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/
2. Hao, K.: I started crying: Inside Timnit Gebru's last days at Google—and what happens next. MIT Technology review. (2020). https://www.technologyreview.com/2020/12/16/1014634/google-ai-ethics-lead-timnit-gebru-tells-story/
3. Johnson, K.: Google targets AI ethics lead Margaret Mitchell after firing Timnit Gebru. VentureBeat (2021). https://venturebeat.com/2021/01/20/google-targets-ai-ethics-lead-margaret-mitchell-after-firing-timnit-gebru/
4. Noble, S.: Algorithms of oppression: how search engines reinforce racism. NYU Press, New York (2018)
5. Devlin, J. and Chang, M.: Open Sourcing BERT: State-of-the-art pre-training for natural language processing (2 November 2018) Posted at https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
6. Bayern, M.: Top 10 US universities that produce the most staff for global tech firms. TechRepublic in CXO, 21 May 21 2020. https://www.techrepublic.com/article/top-10-universities-that-produce-the-most-staff-for-global-tech-firms/
7. Late lessons from early warnings: science, precaution, innovation (chapter 28). EEA Report 1/2013. http://www.eea.europa.eu/publications/late-lessons-2
8. National Whistleblower Center: whistleblowers still the best at detecting fraud. From PricewaterhouseCoopers' 2007 Global Economic Crime Survey. https://www.whistleblowers.org/news/whistleblowers-still-the-best-at-detecting-fraud/
9. Loyens, K., Vandekerckhove, W.: Whistleblowing from an international perspective: a comparative analysis of institutional arrangements. Administrative sciences, MDPI, Published: 5 July 2018. https://www.researchgate.net/publication/326201433_Whistleblowing_from_an_International_Perspective_A_Comparative_Analysis_of_Institutional_Arrangements
10. Benjamin, R.: Race after technology: abolitionist tools for the new jim code. Polity, Cambridge (2019)
11. Boddington, P.: Towards a code of ethics for artificial intelligence. Springer, New York (2017). ISBN 978–3–319–60648–4
12. Bostrom, N.: Superintelligence: paths, dangers, strategies. Oxford University Press, London (2014). ISBN: 9780199678112
13. Buolamwini, J., Gebru, T.: Gender shades project: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, pp. 77–91 (2018). https://www.media.mit.edu/projects/gender-shades/overview/
14. Coeckelbergh, M.: AI ethics. MIT Press, Cambridge (2020) ISBN: 9780262538190
15. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat Mach Intell **1**, 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
16. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. Int. Data Privacy Law **7**(2), 76–99 (2017). https://doi.org/10.1093/idpl/ipx005
17. Wagner, B.: Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? In: Emre, B., Irina, B., Liisa, J.U.A. (Hg.): Being Profiled: Cogitas Ergo Sum. 10 Years of 'Profiling the European Citizen'. Amsterdam University Press, Amsterdam, pp. 84–88 (2018). doi: https://doi.org/10.25969/mediarep/13281
18. Schatsky, D., Katyal V., Iyengar, S. and Chauhan R.: Can AI be ethical? Why enterprises shouldn't wait for AI regulation.

(2019). https://www2.deloitte.com/content/dam/insights/us/articles/4604_S4S-AI-and-ethics/DI_S4S-AI-and-ethics.pdf
19. Karlosson, P., Turner, M. and Gassmann, P.: Succeeding the long-serving legend in the corner office. Leadership. (2019).*
20. Dave, P., Dastin, J.: Google told its scientists to 'strike a positive tone' in AI research – documents. Reuters. (2020). https://www.reuters.com/article/us-alphabet-google-research-focus-idUSKBN28X1CB
21. Ochigame, M.: The invention of "ethical AI'. The Intercept. (2019). https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/
22. Hickok, M.: Lessons learned from AI ethics principles for future actions. AI Ethics (2020). https://doi.org/10.1007/s43681-020-00008-1
23. Hagendorff, T.: The ethics of ai ethics: an evaluation of guidelines. Mind. Mach. **30**, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
24. Haibe-Kains, B., Adam, G.A., Hosny, A., et al.: Transparency and reproducibility in artificial intelligence. Nature **586**, 14–16 (2020). https://doi.org/10.1038/s41586-020-2766-y
25. NeurIPS: Thirty-fourth conference on neural information processing systems. Accessed 12 Jan 2020. https://neurips.cc/Conferences/2020/Schedule?showEvent=16151
26. Abdalla, M., Abdalla, M.: The Grey Hoodie Project: big tobacco, big tech, and the threat on academic integrity. 2020. ArXiv, abs/2009.13676
27. Stahl, B., Chatfield, K., Holter, C., Brem, A.: Ethics in corporate research and development: can responsible research and innovation approaches aid sustainability? J. Clean. Prod. (2019). https://doi.org/10.1016/j.jclepro.2019.118044
28. West, S.M., Whittaker, M., and Crawford, K.: Discriminating systems: gender, race and power in AI. AI Now Institute. (2019). https://ainowinstitute.org/discriminatingsystems.html
29. Raji, I. et al.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. (2020). https://arxiv.org/abs/2001.00973
30. Raji, I., Buolamwini, J.: Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. Assoc. Comput. Mach (2019). https://doi.org/10.1145/3306618.3314244
31. Canca, C.: Operationalizing AI Ethics principles. Commun. ACM **63**(12), 18–21 (2020)
32. Metcalf, J., Moss, E., & Boyd, D.: Owning ethics: corporate logics, Silicon Valley, and the Institutionalization of Ethics. Soc. Res. **86**(2), 449–476 (2019). https://www.muse.jhu.edu/article/732185.
33. Schiff, D., Biddle, J., Borenstein, J., Laas, K.: What's next for ai ethics, policy, and governance? a global overview. Assoc. Comput. Mach (2020). https://doi.org/10.1145/3375627.3375804
34. Eitel-Porter, R.: Beyond the promise: implementing ethical AI. AI Ethics (2020). https://doi.org/10.1007/s43681-020-00011-6
35. MacDougall, D.: Whistleblowing and the Bioethicist's Public Obligations. Camb. Q. Healthc. Ethics **23**(4), 431–442 (2014). https://doi.org/10.1017/S0963180114000103
36. Jobin, A., Lenca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat Mach Intell **1**, 389–399 (2019)
37. Krijger, J. Enter the Metrics: Critical Theory and the Organizational Operationalization of AI Ethics. In press
38. AccessNow: Access Now resigns from the Partnership on AI, 13 October 2020. https://www.accessnow.org/access-now-resignation-partnership-on-ai/
39. Partnership On AI: The Partnership on AI brings together diverse, global voices to realize the promise of artificial intelligence. Accessed 12 Jan 2021. https://www.partnershiponai.org/
40. European Environment Agency (EEA): Late lessons from early warnings: science, precaution, innovation. EEA Report. (2013). https://www.eea.europa.eu/publications/late-lessons-2
41. Gebru, T.: Understanding the limitations of AI: when algorithms fail. Spark + AI Summit (2020). https://databricks.com/session/understanding-the-limitations-of-ai-when-algorithms-fail
42. Gebru, T. *et al.*: Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. (2017) https://www.pnas.org/content/114/50/13108
43. Jo, E. and Gebru, T.: Lessons from archives: strategies for collecting sociocultural data in machine learning. (2020). https://dl.acm.org/doi/abs/https://doi.org/10.1145/3351095.3372829
44. Smith, P., Smith, L.: Artificial intelligence and disability: too much promise, yet too little substance? AI Ethics (2020). https://doi.org/10.1007/s43681-020-00004-5
45. Denton, E. et al.: Image counterfactual sensitivity analysis for detecting unintended bias. (2020). https://arxiv.org/abs/1906.06439
46. Denton, E. et al.: On the dangers of stochastic parrots: Can language models be too big. (2021). https://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf
47. Schrittwieser, J. et al.: MuZero: Mastering Go, chess, shogi and Atari without rules. (2020). https://deepmind.com/blog/article/muzero-mastering-go-chess-shogi-and-atari-without-rules
48. The AlphaFold Team: AlphaFold: A solution to a 50-year-old grand challenge in biology. (2020). https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
49. Breakthrough Technology for the Brain. (2021). https://neuralink.com/
50. Sunyaev, S.: Distributed Ledger technology. (2020). https://link.springer.com/chapter/https://doi.org/10.1007/978-3-030-34957-8_9
51. Kendall, T.: Testimony at the hearing on mainstreaming extremism: social media's role in radicalizing America. Energy and Commerce (116th Congress) and Consumer Protection & Commerce (116th Congress) committees. (2020). https://energycommerce.house.gov/committee-activity/hearings/hearing-on-mainstreaming-extremism-social-media-s-role-in-radicalizing
52. Big Tobacco Guilty of Lying to the Public. Tobacco stops with me. https://stopswithme.com/exposing-big-tobacco/big-tobacco-found-guilty/
53. Rugaber, W: Industry resists car-safety costs. Special to The New York Times (1975).https://www.nytimes.com/1975/04/06/archives/industry-resists-carsafety-costs-companies-feel-consumers-will.html
54. Nader, R.: Unsafe at any speed: the designed-in dangers of the American automobile. Am. J. Publ. Health **101**(2), 254–256 (2011). https://doi.org/10.2105/ajph.101.2.254
55. Digital Technology and the Resurrection of Trust: House of Lords Select Committee on Democracy and Digital Technologies. Report of Session 2019–21. HL Paper 77. (2020). https://committees.parliament.uk/publications/1634/documents/17731/default/
56. Schwab, K.: 'This is bigger than just Timnit': How Google tried to silence a critic and ignited a movement. FastCompany. (2021). https://www.fastcompany.com/90608471/timnit-gebru-google-ai-ethics-equitable-tech-movement
57. Benjamins, R.X., Barbado, A., and Sierra, D.: Responsible AI by Design in Practice. In: Proceedings of the Human-Centered AI: Trustworthiness of AI Models & Data (HAI) track at AAAI Fall Symposium, DC, 2019