CONSTANTIN VICĂ, CRISTINA VOINEA,
RADU USZKAI

The emperor is naked: Moral diplomacies and the ethics of AI

With AI permeating our lives, there is widespread concern regarding the proper framework needed to morally assess and regulate it. This has given rise to many attempts to devise ethical guidelines that infuse guidance for both AI development and deployment. Our main concern is that, instead of a genuine ethical interest for AI, we are witnessing moral diplomacies resulting in moral bureaucracies battling for moral supremacy and political domination. After providing a short overview of what we term 'ethics washing' in the AI industry, we analyze the 2021 UNESCO Intergovernmental Meeting of Experts (Category II) tasked with drafting the Recommendation on the Ethics of Artificial Intelligence and show why the term 'moral diplomacy' is better suited to explain what is happening in the field of the ethics of AI. Our paper ends with some general considerations regarding the future of the ethics of AI.

Keywords: *moral diplomacies, moral bureaucracy, AI ethics, AI guidelines, ethics washing*

Acknowledgement

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS -UEFISCDI, project number PN-III-P1-1.1-TE-2019-1765, within PNCDIIII, awarded for the research project *Collective moral responsibility: from organizations to artificial systems. Re-assessing the Aristotelian framework*, implemented within CCEA & ICUB, University of Bucharest (2021-2022).

Author Information

Constantin Vică, Faculty of Philosophy, University of Bucharest
https://orcid.org/0000-0001-8975-8827
Cristina Voinea, Bucharest University of Economic Studies
https://orcid.org/0000-0003-4654-0697
Radu Uszkai, Bucharest University of Economic Studies
https://orcid.org/0000-0001-5250-8015

How to cite this article:

Vică, Constantin, Cristina Voinea, Radu Uszkai. "The emperor is naked: Moral diplomacies and the ethics of AI."
Információs Társadalom XXI, no. 2 (2021): 83–96.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.6

1. Introduction

Artificial intelligence (AI) is a shining star within the technology world. All other technological innovations and artefacts pale in comparison with what AI, in all its shapes and sizes, promises to offer. However, as the saying goes, all that glitters is not gold. AI technologies have pushed the significance of dual-use to the extreme: whether we think about autonomous weapons, facial recognition technologies or already mundane decision-making software, all of these applications can be used for both good and bad purposes. For example, decision-making algorithms can improve efficiency, but they can also reinforce racial prejudices and biases as they may discriminate based on race or gender (Buolamwini and Gebru 2018; Noble 2018). Other AI systems, such as scoring systems, identify and exploit weaknesses that individuals may not be aware of themselves (Citron and Pasquale 2014). And while discrimination, manipulation or exploitation have plagued societies since the dawn of civilization, unlike with human decision-making, AI systems can operate at scale, instantly and automatically, with the potential to affect people in the flash of a second, "at orders of magnitude and at speeds not previously possible" (Yeung, Howes and Pogrebna 2019). It is precisely these actual and potential harms that AI could create that have driven the massive interest in the ethics of AI.

In this paper, we explore the implications and consequences of the particular interests of both private companies and states alike for the development of ethical guidelines for AI systems. In the first section, we look at some critiques of private companies' focus on the development of ethical codes of conduct and guidelines for ethical AI. We show that most researchers tend to focus on the problem of 'ethics washing', which is the superficial and even hypocritical use of ethics for the avoidance of state regulation. The criticism of companies' attempts to self-regulate is based on the belief that they will always strive to advance their own interests, thus their efforts of devising ethical or responsible AI systems will not do away with the wider problems generated by the societal deployment of these technologies. However, a hidden presupposition behind these types of analyses is that if private companies should not be left alone to their own devices, then states should take the lead in the efforts to advance ethics in the field of AI. In the following section, we show that, in practice, states don't fare too well in this domain either. We take as a case study UNESCO's attempt to create yet more guidelines for ethical AI, in order to show that both transnational organizations and states alike use ethics as a locus of power. We advance the term 'moral diplomacy' to describe the strategy of using the language of morality, by transnational organizations, states and the industry alike, to protect and advance forms of technology that can advance certain economic and political interests. In the concluding remarks, we claim that the fight over 'AI ethics' is actually a political fight and that the ethical guidelines and regulations for AI advanced by 'moral diplomacies' are just a way of signalling allegiance to certain ethical values and principles, without actually moving towards their accomplishment.

2. From ethics washing to the bureaucratization of ethics

While the ethical implications of AI have been addressed since the 1960s, the emergence of machine learning and neural networks has brought ethical debates in to the mainstream (Morley et al. 2019). ‘AI ethics’ is now sort of a buzzword in the field, as it is employed to name and describe a whole array of moral, legal, societal and political concerns associated with the development and implementation of AI technologies. One of the most frequently employed tools that is believed could help resolve the ethical issues generated by AI are documents containing ethical principles, frameworks, checklists and guidelines to aid the development and implementation of AI technologies. These documents are considered a universal panacea for the potential harms generated during technological development and implementation, a fact shown by the diversity and multiplicity of organizations that have rushed to issue such documents, from industry, to governments, transnational organizations, academia and NGOs. An exhaustive list of all these documents and organizations would probably take the whole space of this paper, so here we settle with mentioning just a few of them: IEEE’s (2019) *Ethically Aligned Design “Vision”*, *Artificial Intelligence at Google* (2018) manifesto, OpenAI, Partnership on AI or The Foundation for Responsible Robotics. Jobin et al. (2019) identified no less than 84 documents of a non-legal nature (research and position papers excluded) expounding ethical principles and guidelines for AI. Most of these documents are issued by private companies, followed by governmental agencies, while academic institutions, supposedly the only impartial and objective organizations, are the last issuers of such recommendations and guidelines (AI Ethics Lab 2020; AlgorithmWatch 2020).

Although the attention paid to ethics in AI development and deployment is a heartening development, the focus on ethical guidelines is not without its critics. One of the first problems identified by the critics is that most of these documents are principle-based, embracing a deontological approach (Mittelstadt 2019). Principles are highly abstract standards for good, but they tend to be vague as their application is most of the time context sensitive. As a consequence, principle-based AI guidelines have been criticized for not being sufficiently action-guiding (Hagendorff 2020b; Héder 2020). This means that it is not clear to AI practitioners how to put these principles into practice, as principles, by themselves, do not play a role in informing and training the moral reasoning needed for ethical behaviour in a practical context (Greene, Hoffmann and Stark 2019). This is further proven by the fact that, despite the richness in ethical guidelines, 79% of tech workers report that they would like more practical, down-to-earth instructions on how to deal with and address ethical problems in technology development (Miller and Coldicott 2019).

The ineffectiveness of professional codes of conduct or of any sort of guidelines for the development of responsible AI is further complicated by the fact that AI systems can be used in a wide range of domains, from medicine to warfare and many others. Further, AI developers do not have a common background, as they come at AI from various domains and will be specialized in different disciplines, which also means that they might have different moral obligations to attend to (Filipović, Koska and Paganini 2018). However, ethical guidelines tend to reduce AI developers to a single expertise, which cannot but obscure the complexity of reality (Mittelstadt 2019). Moreover, any sort of deviation from these principles would be hard to notice and also difficult to punish, as these documents lack enforcement mechanisms (Hagendorff 2020a; LaCroix and Mohseni 2020).

Another contentious issue connected to these ethical guidelines is the lack of diversity of their creators. In his analysis of 22 AI ethics guidelines, Hagendorff (2020b) shows that the ratio of female to male authors is 31.3%, and makes an interesting observation that those reports authored primarily by men tended to focus on particular issues, such as privacy or transparency, ignoring the fact that when AI systems are deployed, they become embedded in complex sociotechnical systems. This shows that male-dominated reports tend to oversimplify the problems these technologies give rise to when they complement or even substitute human decision-making, ignoring important issues such as welfare, fairness or even ecological concerns (Hagendorff 2020b). What is more, Jobin et al. (2019), in their analysis of the corpus of the principles and guidelines on ethical AI, noticed an underrepresentation of developing regions, such as Africa, Central and South America and Central Asia, which of course denotes an existing global power imbalance that it seems is even perpetuated in AI ethics debates. This raises questions of global fairness, but it also denotes a sort of technological determinism implicit in most documents. It is almost as if humans can only react to these technologies as if they are a force that we cannot shape (Greene, Hoffmann and Stark 2019). Further, most documents have as their locus design processes, mostly ignoring business or political decisions, revenue models or the incentive mechanisms that after all shape design processes (Yeung, Howes and Pogrebna 2019).

If the lack of specificity and diversity were the sole issues with these ethical guidelines for the development and deployment of AI, then there would be no significant reasons to worry. After all, these are problems that could, in principle, be solved by more careful deliberation and consideration of the purposes and application of ethical guidelines/codes of conduct. Another, more important worry, though, is that these high-level principles and documents are used as a façade by the industry, and essentially as a ploy to delay or plainly avoid policy-maker's reasons to pursue regulation. To put it more simply, the underlying idea in almost all of these documents is that states' role in regulating AI technologies can be sidelined, while the role of the private sector should be overly-emphasized (Wagner 2018). In 2019, the term 'ethics washing' was first used by the philosopher Thomas Metzinger to describe the instrumentalization of ethics by industry, in his critique of the European Commission Ethics Guidelines for Trustworthy AI (Metzinger 2019). Responsible for the creation of this document was the 52-member High-Level Expert Group on Artificial Intelligence (HLEG AI), which was heavily dominated by industry, with only four ethicists part of the team. Metzinger complains that the guidelines issued by HLEG AI are "lukewarm, short-sighted and deliberately vague" precisely because ethics is instrumentalized in order to "distract the public and to prevent or at least delay effective regulation and policy-making" (Metzinger 2019).

The inspiration for this term ethics washing comes from the already popular term 'greenwashing'. The suffix '-washing' is used to denote a gap between the behaviour of a business or government and how that behaviour is framed or communicated to the public (Peukert and Kloker 2020). While greenwashing refers to the discrepancy between the claims companies make about the environmental impact of their products/services and their actual environmental impact (Voinea and Uszkai 2020), ethics washing denotes the proclaimed adherence to ethical standards by AI companies in order to escape regulation and to reassure customers and other stakeholders of their ethical commitment (Bietti 2020; Wagner 2016; Peukert and Kloker 2020; Reséguier and Rodrigues 2020). Besides in the creation of AI working groups meant to issue guidelines for ethical AI, ethics washing is also manifested in ethics partnerships for AI, such as in the employment by industry of in-house philosophers and ethicists with little or no influence on design processes or business operations (Bietti 2020), and also in the funding by Big Tech of academic work on responsible or ethical AI, which is really meant to obscure problems regarding business practices or the political implications of AI systems (Abdalla and Abdalla 2021; Ebell et al. 2021).

The use and abuse of ethics within the technology world seems to be a strategy employed by various stakeholders in order to create the impression, for both the public and governments, that internal self-regulation by science and industry is more than enough for dealing with the risks raised by AI systems (Bietti 2020). In a paradoxical turn of events, ethics is now used to protect and foster the status quo, while eliminating the possibility of moral progress in the technological world. Many important and stringent ethical implications of AI technologies, such as the social and political impacts of algorithmic decision-making, the environmental implications of data processing for AI, and the rise of fake news/propaganda/deep-fakes, as well as the private funding of public research institutions in the field of AI remain virtually unaddressed within these documents (Hagendorff 2020b).

In what follows, we claim that ethics washing is not the most appropriate way to describe the instrumentalization of ethics in the technology world, because it tends to frame the avoidance of regulation of technology companies by public authorities as something that is bad in itself. But the question of whether governments are better placed to regulate complex, constantly evolving and changing technologies, such as ML-based AI, remains unaddressed. We advance the term 'moral diplomacy' to describe the strategy of using the language of morality, by both transnational organizations and the industry alike, to protect and advance forms of technology that can advance certain economic and political interests. Just as the moral diplomacy conceived by US President Woodrow Wilson was an instrument of fighting back against governments that opposed or were hostile to American interests, so the moral diplomacies in today's AI landscape are a way of advancing political and economic interests and of nipping in the bud discussions addressing important questions about power arrangements. In the following section, we show what moral diplomacies may consist of by analyzing the UNESCO Intergovernmental Meeting of Experts (Category II) tasked with drafting the Recommendation on the Ethics of Artificial Intelligence.

3. The birth of moral diplomacy and AI governance

Despite the views of Immanuel Kant (1998), in day-to-day life, ethics is seldom "pure", that is based solely on supreme or ultimate moral principles. When it comes to applying ethics in practice, like in the development of ethical guidelines for AI systems, ethics is surrounded by many other 'vocabularies' and intellectual disciplines: law and legal thinking (especially human rights), economic and institutional approaches, but also political stakes or social opportunities, etc. The main peril for ethics in the highly dynamic landscape of AI is for it to become just a pretext or a decorative, floral *adagio* in attempts to protect and entrench the *status quo*. When ethics becomes mere etiquette, it fails to provide deliberative mechanisms, sound judgment and true answers. Other risks are not to be neglected either: ethics could become an instrument of struggle or persuasion, a motive for negotiation (that involves *trading* and not *pondering* values or principles) or even a way of 'washing' the image of companies. To put it simply, ethics is in danger of becoming a (cultural or even technological) instrument of domination.

What happens *with* AI ethics and the attempts to codify it (in the form of recommendations or White Papers) is the continuation of a trend that started in early modernity. In search of impartiality, ethics is de-personalized, becoming an art of legalization, architectonics or building systems (Iftode 2021). Moreover, ethicists are beginning to lose sight of a fundamental problem in ethics, that is, moral motivation (Iftode 2021). Moral motivation cannot be merely extrinsic, it cannot lie only in the power of a law, a precept, or of any recommendation, no matter how convincing it is. It should be clear for everyone that governing AI systems, for their lifetime cycle, through ethical codes and guidelines, or recommendations is not a solution, but is increasingly becoming part of the problem. AI is not like the commons – be it pastures, rivers or Wikipedia – for which there are models of good collective governance (Ostrom and Hess 2007). AI systems are not common resources (although maybe data should be), and perhaps that is why the open-source development model has not caught on in the AI research world.

We call 'moral diplomacies' the widespread arrangements of negotiating and gaining *consensus* on the moral guidelines for AI development. Until now, there have been at least three notable productions of moral diplomacies: the OECD Recommendation of principles (adopted on May 22, 2019), the EU guidelines, and the in-the-making UNESCO Recommendation. The outputs of moral diplomacies are documents, in a word-based format, which necessarily implies the emergence of moral bureaucracies capable of interpreting and making decisions on their basis. This is a mechanism similar to academic or medical ethical committees, or Institutional Review Boards (IRBs), the institutions putting ethical codes into practice (Molina and Borgatti 2019). In short, ethical AI governance, transcribed in codes or recommendations, is a product of moral diplomacies, further creating moral bureaucracies.

In what follows, we focus on UNESCO's approach to AI ethics, mainly because it is one of the most transparent and open to inquiry[1] cases of moral diplomacy, allowing

[1] This goes hand in hand with the subjective reason for choosing UNESCO: one of the authors was an expert participating in the discussions.

a detailed analysis. Not only was the draft Recommendation made public (UNESCO 2021a), but also the Intergovernmental Meeting of Experts (Category II) tasked with creating the Draft Recommendation on the Ethics of Artificial Intelligence was livestreamed and kept online afterwards (UNESCO 2021b). It is also important to stress that this forthcoming Recommendation is non-binding, i.e. it has no legal effects and creates no obligations (compared to a Convention, which should be instilled in national legislations), and it will not come into effect before being accepted by member states in another high-level meeting, namely the UNESCO General Conference. Before the Intergovernmental Meeting, there was an arduous process of drafting the Recommendation, prepared by the Ad Hoc Expert Group (AHEG) based on wide multistakeholder consultations. The Recommendation included a preamble and 141 articles structured around the aims, objectives, values, principles and areas of policy action (UNESCO 2021). It was accompanied by a preliminary study and a final report. Also, before the meeting, the member states were invited to send their comments and amendments, which in turn produced a huge document of almost 1000 pages. So, the amount of work and the outputs was highly impressive. From this point on, the deliberation began in earnest. Keep in mind here that our short analysis is limited to the first session of the Intergovernmental Meeting of Experts (26–30 April 2021).

This debate is representative for the making of public AI ethics. If we take the ideal model of discourse ethics (Habermas 1990; Bohman and Rehg 2017) as a frame of reference, we can see that not all of its “pragmatic presuppositions” have been fulfilled. First of all, (1) we need to use the linguistic expressions in the same way in order to ensure we have the same meanings in play; then, (2) none of the relevant arguments can be ruled out. Third, we must take into consideration (3) only the strength of the arguments, and not their rhetorical power of persuasion. For things to work, (4) all the participants must be motivated to find the best argument. Last but not least, (5) no-one should be excluded. The result of the deliberative process should be the intellectual empowerment of the participants, and its foundation lies precisely in the equal respect accorded to everybody involved. Undoubtedly, equal respect was given to all the participants who were able to intervene and propose amendments to the articles of the Recommendation. Condition (5) was met, in that no-one present was excluded. It should be noted, however, that not all states were represented, with some having only observers there (such as the USA, which had withdrawn from UNESCO) or were altogether absent.[2] The first condition was impossible to fulfil in practice; for example, participants had different meanings for some of the more contentious terms, such as ‘gender equality’ or ‘universal’, meanings that did not necessarily converge. Further, some arguments were ruled out – which contradicts condition (2) – only because of different experts’ rhetorical power of persuasion – thus going against condition (3). Further, the experts were not limited to ethicists and AI researchers, many of them were human rights lawyers and activists, or diplomats, appointed by their states to advance specific values and principles resonating with their own foreign

[2] A novelty of this meeting was its online format: it took place on Zoom.

and domestic policies. What was really problematic, though, was the unfulfillment of condition (4): here, the aim was not to find the best arguments, but to block or support various positions without appealing to moral grounding, rather to political expediency.

The specific procedure rules for UNESCO clearly stipulate that discussions on the Recommendation draft may advance by consensus. And here resides the first difference between the philosophical and the diplomatic employment of ethics. Ethical debate cannot have either as a mere goal, or as a method, consensus at any price. Consensus could be the goal of institutions in gaining uniform practices, but philosophical grounding is bound to the strength of the argument and truth finding. Undoubtedly, in a pragmatic sense, reaching agreement is important, but not at the price of distorting its foundation. During this first round of debates, one key question was to find the *sources of normativity* for the Recommendation. Here, the divide was apparent from the beginning (especially during the April 28 meeting): some state representatives insisted on human rights law as having priority and should be the only universal, normative source of the document (UNESCO 2021b). Others insisted on ethics and its particular contribution as an extension into areas that human rights law cannot cover; as an answer to the focus on ethics, some others decried this as 'ethics washing'. One of the participants (observer) said: "The language of ethics has the merit of shedding light on the blind spots in current (positive) international law. At present, the Recommendation has nothing to add to the current legal framework." Also, one ethicist bluntly expressed his opinion in the chat box by stating: "I understand that ethics is out of the scope of this discussion" and then adding a touching quote from Vladimir Jankélévitch, "Evil is the disjunction of virtues, it is to have a virtue without the others". This reaction says a lot about the actual divorce between legalistic thinking and ethical deliberation and, even more, about the unrealistic expectation that an ethical code or recommendation will make AI systems virtuous. Indeed, the delegates spent almost a whole day of the session rejecting an amendment on the role of international law which, in the end, was made less prominent within the Recommendation. We do not wish to comment on whether this is good or bad, right or wrong, but we want in fact to stress that the focal points of the debate did not seem to have in view the ethics of AI, but rather the concealment of the political interests that would like to instrumentalize AI. Article 11 of the draft, "While all the values and principles outlined below are desirable per se, in any practical context there are inevitable trade-offs among them, requiring complex choices to be made about contextual prioritization, without compromising other principles or values in the process, especially human rights and fundamental freedoms" (UNESCO 2021, 7), was the biggest bone of contention. The prolonged debate around it, taking up nearly an entire day, is paradigmatic of the whole context of drafting ethical principles and norms for AI systems. Because the locution 'trade-off' has different meanings within different domains of discourse, the debates regarding what it refers to more precisely, and what ethical values and principles should be prioritized in case of a conflict, almost blocked any advance in reaching agreement.

The inherent conflict between states has moved into the realm of AI ethics, as another attempt to move from the power of arms to the power of speech (which is a fundamental way of preserving peace through culturalization). AI ethics is a territory unclear to many, ideal and conceptual, but with immense material implications. In this game, the clash between lawyers and ethicists is obvious. And even more obvious is the struggle between the 'old world' of human rights, and the 'new world' prone to use AI in governing populations. It parallels the symbolic struggle between universalists (the Western World) and generalists (Iran, China or Venezuela) over the nature of human rights. All this is part of an 'ethical arms race' between organizations (international or industry alike) for exerting influence upon the future of AI development. In all this context, ethics in its practical or applied exercise becomes the loser, the abandoned puppet. As long as these kinds of documents are non-binding, the effort seems directed towards something different from the red-lines or the way AI should be governed for the common good. The incompatibility between ethics, a pluralistic and revisable system (or even fully particularistic sometimes), with its trade-offs, limitations and balancing, and international law, with its positive, rigid foundation, is hard to overcome. This adds to the main issue, namely that when consensual methods are applied in ethics, the risk is that the achieved compromise will totally reduce the normative power of ethical guidelines.

Almost everything that is ethically 'revolutionary' in this kind of document has been or could be eliminated by 'diplomatic games'. For philosophers and ethicists, it is frustrating to see that several conditions of discourse (or argumentation) ethics are not fulfilled and, even worse, that deliberation becomes bartering. Ethics was seen as part of the art of politics by Aristotle. For international organizations, ethics has rather assumed the role of a shield against recognizing the political nature of the creating institutions. For example, in the UNESCO document, there are no remarks about power and the power relationships built around AI. But power asymmetries are real and actual. Moreover, top-down approaches, based on human rights normativity, are necessary, but not sufficient alone.

4. Concluding remarks: Regulating AI, a catch-22 situation?

The presupposition behind criticisms of companies' capacities to self-regulate through ethical guidelines is that states, especially democratic ones, are better suited to take the lead and to impose clear red-lines concerning the development and deployment of AI systems. In the above discourse, we showed that states don't fare too well either in this domain. The ethics of AI, as it is approached today by industry or transnational organizations and states, is yet another proxy for advancing various types of interests – be they financial in the case of private companies, or political in the case of states. This is another example of the fact that technologies are not mere neutral functional tools, but are also ways of doing politics by other means, as Winner argues (Winner 1986). Technologies are political because they are shaped by human choices and institutional structures, and in their turn, they shape the way things are done in a society. The politics of AI systems lies in the fact that such systems can change the distribution of power at a societal level, empowering some, while making others even more vulnerable than before (Voinea 2016). Currently, the 'fight' over AI ethics is actually a fight over the specific forms of power and authority that these technologies should incorporate.

While not as absurd and paradoxical as Yossarian's conundrums from Joseph Heller's famous 1961 novel *Catch-22*, the ethics of AI seems to be in a catch-22 situation. On the one hand, the recent push for the industry's self-regulation has proven to be unsuccessful. Companies have shown only an instrumental interest in the tools that normative and applied ethics can bring to the table for regulating AI. Naturally, one might think that the solution to the drawbacks of this strategy might be to bring states and international organizations in to fill in the gaps, like most economists tend to think that we should do when we face a market failure.

As our analysis of the UNESCO Intergovernmental Meeting of Experts tasked with drafting the Recommendation on the Ethics of AI shows, this heuristic approach is not useful in our case. Our main claim is that, at its core, the issue lies with the fact that the ethical guidelines and regulations for AI are advanced by what we term 'moral diplomacies', which are employed by both private (i.e. the dominant companies from the industry) and public organizations (i.e. states and other international organizations) for elevating their status by the use of 'moral talk' or, as Tosi and Warmke (2020) put it, for grandstanding purposes.

Whether it's for avoiding more robust regulation and attracting better employees, like it would be in the case of a company like Google (Voinea and Uszkai 2020), or for politicians to signal to the electorate that they care about Responsible AI (post-industrial democracies) or to make their opposition to Western democracies and their WEIRD morality (Haidt 2012) internationally known, it has become clear that we cannot solve a political problem with ethical ramifications (the regulation of AI) just by simply drafting codes of ethics and establishing moral bureaucracies. Even if we were to leave aside the classical criticism of bureaucracies and bureaucrats as being simply budget maximizers (Niskanen 1971; 1994), an opaque ethical infrastructure that does not contribute to the development of moral and intellectual virtues for the individuals who actually work with AI (Constantinescu et al. 2021) would be nothing more than a waste of both public and private resources, and with potentially deleterious consequences.

This quasi-pessimistic outlook on the future of AI ethics can be supplemented by an even further troubling implication for ethicists who want to have an impact outside just academia. Some ethicists might wish to shape the outlook of the industry on AI by seeking employment at Google or other major players in the industry. For others, the option of ensuring ethical checks and balances is part of public AI moral bureaucracies. Our claim is that any ethicists who might strive to advance an unbiased agenda for ethical AI and at least aim to marginally improve the current *status quo* of AI ethics will probably face what Walzer famously labelled as "the problem of dirty hands" (Walzer 1973). For example, an ethicist working for Google might have to accept some privacy intrusions for profit-maximizing purposes in order to push for a more robust concern of the company for eliminating unfair biases in the ways in which the company processes data, for instance. Similarly, working as an AI moral diplomat for a Western democracy might mean that a person would need to sacrifice some of their principles either due to the electoral interests of their employer (i.e. the Government and/or political party in power) or because intercultural negotiations might entail an unsettling balancing of human rights in order to push an agenda that could be acceptable for countries with a different moral *weltanschauung*, i.e. world view. The only question that remains, then, is what is the acceptable threshold after which compromises with both industry and states or international organizations alike becomes morally unacceptable.

References

Abdalla, Mohamed, and Moustafa Abdalla. “The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity.” *ArXiv:2009.13676 [Cs]* (April 2021). https://doi.org/10.1145/3461702.3462563.

AI Ethics Lab. “Tool: The Box.” *Toolbox: Dynamics of AI Principles*, June 2020, https://aiethicslab.com/the-box/.

AlgorithmWatch.“AI Ethics Guidelines Global Inventory by AlgorithmWatch.” Retrieved May 8 2021. https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/.

Bietti, Elettra. “From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy.” In *FAT* ’20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–19. New York: Association for Computing Machinery, 2020. https://doi.org/10.1145/3351095.3372860.

Bohman, James, and William Rehg. “Jürgen Habermas.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Metaphysics Research Lab, Stanford University, Fall 2017. https://plato.stanford.edu/archives/fall2017/entries/habermas/.

Buolamwini, Joy, and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of Machine Learning Research. Conference on Fairness, Accountability and Transparency*, 81 (2018): 77–91.

Miller, Catherine, and Rachel Coldicott. “People, Power and Technology: The Tech Workers’ View.” Retrieved June 8 2021.
https://doteveryone.org.uk/report/workersview/.

Citron, Danielle Keats, and Frank Pasquale. “The Scored Society: Due Process for Automated Predictions”. *Washington Law Review* 89, no. 1 (January 2014): 1–33.

Constantinescu, Mihaela, Cristina Voinea, Radu Uszkai, and Constantin Vică. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context." Unpublished manuscript, April 2021. https://www.researchgate.net/publication/352519451_Understanding_responsibility_in_Responsible_AI_Dianoetic_virtues_and_the_hard_problem_of_context

Ebell, Christoph, Ricardo Baeza-Yates, Richard Benjamins, Hengjin Cai, Mark Coeckelbergh, Tania Duarte, Merve Hickok, Aurelie Jacquet, Angela Kim, Joris Krijger, John MacIntyre, Piyush Madhamshettiwar, Lauren Maffeo, Jeanna Matthews, Larry Medsker, Peter Smith, and Savannah Thais. "Towards Intellectual Freedom in an AI Ethics Global Community." *AI and Ethics* 1, no.2 (May 2021): 131–38. https://doi.org/10.1007/s43681-021-00052-5.

Filipović, Alexander, Christopher Koska, and Claudia Paganini. "Developing a Professional Ethics for Algorithmists." *Working Paper. Bertelsmann Stiftung* 2018. Retrieved May 8 2021. https://www.bertelsmann-stiftung.de/en/publications/publication/did/developing-a-professional-ethics-for-algorithmists.

Google. "Artificial intelligence at Google: Our principles." 2018. Retrieved May 8, 2021. https://ai.google/principles/.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, edited by Tung X. Bui, 2122–2131. Honolulu: HICSS, 2019.

Habermas, Jürgen. *Moral Consciousness and Communicative Action*. Cambridge (MA): MIT Press, 1990.

Hagendorff, Thilo. "AI Virtues–The Missing Link in Putting AI Ethics into Practice." *ArXiv Preprint ArXiv:2011.12750,* 2020a.

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30, no.1 (March 2020b): 99–120.

Haidt, Jonathan. *The Righteous Mind. Why Good People are Divided by Politics and Religion*, London: Penguin, 2012.

Héder Mihály. "A criticism of AI ethics guidelines." *Információs Társadalom* XX, no. 4 (2020): 57–73.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. "Ethically Aligned De-sign: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition." 2019. Retrieved May 8, 2021. https://standards.ieee.org/content/ieee-stand-ards/en/industry-connections/ecautonomous-systems.html

Iftode, Cristian. *Viața Bună. O Introducere În Etică*. București: Trei, 2021.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1, no. 9 (2019): 389–99.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press, 1998.

LaCroix, Travis, and Aydin Mohseni. "The Tragedy of the AI Commons." *ArXiv Preprint ArXiv:2006.05203* (2020).

Metzinger, Thomas. "Ethics washing made in Europe." *Der Tagesspiegel*. August 4, 2019. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html.

Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." *SSRN Scholarly Paper* ID 3391293 (2019). https://doi.org/10.2139/ssrn.3391293.

Molina, José Luis, and Stephen P. Borgatti. "Moral Bureaucracies and Social Network Research." *Social Networks*, (November 2019). https://doi.org/10.1016/j.socnet.2019.11.001.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. "From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices." *ArXiv Preprint ArXiv:1905.06876* (2019).

Niskanen, William. A. *Bureaucracy and Public Economics*, Aldershot: Edward Elgar, 1994.

Niskanen, William A. Bureaucracy and Representative Government, Chicago: Aldine Atherton, 1971.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press, 2018.

OECD. "Recommendation of the Council on Artificial Intelligence." *OECD Legal Instruments*, 2019. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

Ostrom, Elinor, and Charlotte Hess. "A Framework for Analyzing the Knowledge Commons." In *Understanding Knowledge as a Commons: From Theory to Practice*, edited by Charlotte Hess and Elinor Ostrom, 41–81. Cambridge (MA): MIT Press, 2007.

Peukert, Christian, and Simon Kloker. "Trustworthy AI: How Ethics Washing Undermines Consumer Trust." In *Proceedings of the 15th International Conference on Wirtschaftsinformatik, Potsdam*, 2020. https://Doi. Org/10.30844/Wi_2020_j11-Peukert.

Rességuier, Anaïs, and Rowena Rodrigues. "AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics." *Big Data & Society* 7, no. 2 (July-December 2020): 1-5. https://doi.org/10.1177/2053951720942541.

Tosi, Justin, and Brandon Warmke. *Grandstanding. The use and abuse of moral talk*. New York: Oxford University Press, 2020.

UNESCO. "Draft Text of the Recommendation on the Ethics of Artificial Intelligence." *SHS/IGM-AIETHICS/2021/APR/4*. Paris: UNESCO, 2021a. https://unesdoc.unesco.org/ark:/48223/pf0000376713.

UNESCO. "Intergovernmental Meeting of Experts (Category II) related to a Draft Recommendation on the Ethics of Artificial Intelligence." 2021b. http://webcast.unesco.org/events/2021-04-REC-Ethics-of-AI/.

Voinea, Cristina, and Radu Uszkai. "Do Companies Engage in Moral Grandstanding?" In *Proceedings of the International Management Conference*, edited by Ion Popa, Cosmin Dobrin, Carmen Nadia Ciocoiu, 1033–1039. Bucharest: ASE University Press, 2020.

Voinea, Cristina. "Governance without Governors: Politics through Algorithms and Big Data." *Revista de Filosofie*, LXIII, no. 6 (2016): 583–595.

Walzer, Michael. "Political Action: The Problem of Dirty Hands." *Philosophy & Public Affairs* 2, no. 2 (Winter, 1973): 160–180.

Wagner, Ben. "Algorithmic Regulation and the Global Default: Shifting Norms in Internet Technology." *Etikk i Praksis - Nordic Journal of Applied Ethics* 10, no. 1 (2016): 5–13.

Wagner, Ben. "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping." In *Being Profiling. Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, edited by Emre Bayamlıoğlu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt, 1–7. Amsterdam: Amsterdam University Press, 2018.

Winner, Langdon. *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. Chicago: University of Chicago Press, 1986.

Yeung, Karen, Andrew Howes, and Ganna Pogrebna. “AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing.” *SSRN Scholarly Paper* ID 3435011 (2019). https://doi.org/10.2139/ssrn.3435011.