Philosophy & Technology (2022) 35: 45
https://doi.org/10.1007/s13347-022-00538-y

COMMENTARY

In Defence of Principlism in AI Ethics and Governance

Elizabeth Seger[1,2]

Received: 8 April 2022 / Accepted: 18 April 2022 / Published online: 28 April 2022

Abstract
It is widely acknowledged that high-level AI principles are difficult to translate into practices via explicit rules and design guidelines. Consequently, many AI research and development groups that claim to adopt ethics principles have been accused of unwarranted "ethics washing". Accordingly, there remains a question as to if and how high-level principles should be expected to influence the development of safe and beneficial AI. In this short commentary I discuss two roles high-level principles might play in AI ethics and governance. The first and most often discussed "start-point" function quickly succumbs to the complaints outlined above. I suggest, however, that a second "cultural influence" function is where the primary value of high-level principles lies.

Keywords Principlism · AI governance · Biomedical ethics · Cultural norms · Ethics washing

1 Introduction

Recent reports estimate that there are at least 70 publicly available sets of ethical AI principles proposed by governments, AI companies, and independent AI ethics initiatives (Floridi et al., 2018; Jobin et al., 2019). However, it is also widely acknowledged that such high-level principles are difficult to translate into practices via explicit rules and design guidelines (Morley et al., 2020; Peters et al., 2020).

Consequently, many AI research and development groups that claim to adopt ethics principles have been accused of unwarranted virtue-signaling or "ethics washing" (Nemitz, 2018). According to these critics, ethics principles not demonstrably implemented into AI research and development practices serve as little more than smoke and mirrors that project an unwarranted image of trustworthiness to

✉ Elizabeth Seger
eas97@cam.ac.uk

[1] Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK

[2] Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

consumers, and which provide policymakers with reasons not to pursue enforceable regulation (Calo, 2017; Mittelstadt, 2019; Whittlestone et al. (2019b)).

Accordingly, there remains a question as to if and how high-level principles should be expected to influence the development of safe and beneficial AI. In this paper, I agree that the translation of high-level principles into practice poses a significant challenge; however, we should not be so quick to dismiss principles as a tool in the AI governance repertoire. In addition to serving as a first step toward articulating more specific ground-level rules and practices, high-level principles also play a key role in promoting and maintaining cultural values and behavioral norms in professional communities which are conducive to the uptake and increased efficacy of more explicit rules and requirement. In virtue of this second role of AI ethics principles, I argue for a dual approach to AI governance in which principlism is supplementary to external policy and recognized as instrumental to the successful implementation of extrinsic rules and regulations.

2 What Is the Point of a Principle?

To understand the role of principles in AI governance, medical ethics is a good starting point.[1] Consider the medical mantra "do no harm", the Hippocratic underpinning of nonmaleficence, one of four biomedical principles put forth by Beauchamp and Childress (1979). "Do not harm" or nonmaleficence is a central tenet of the medical profession, but it is not clear why such a principle needs to be stated. On the one hand, it seems obvious that a physician should not intentionally hurt a patient, but, on the other hand, it is too broad to give specific guidance. For example, there is much debate as to whether physician-assisted suicide aligns with or runs counter to the nonmaleficence principle (Shibata, 2017). So, why state a principle if it is either too obvious or too broad?

2.1 Principles Are the First Step Toward Articulating Ethical Practice

The most obvious answer, and one that few would dispute, is that principles provide a starting point for articulating more precise rules and requirements for ethical practice. For instance, simply stating "do no harm" raises questions such as what does "do no harm" mean in the context of physician-assisted suicide, abortion, or a painful treatment regime? Relevant debates then ensue.

Within the framework of AI ethics and governance literature, principles are primarily assumed to play such a "start-point" function. Due to the nature of their generality, principles cannot be expected to provide specific ground-level guidance in and of themselves. However, they do perform a useful function in categorizing ethical issues for further research, and as such, provide a point of departure for thinking

[1] Others have made the same observation (Cave et al., 2021; Floridi et al., 2018; Mittelstadt 2019; Whittlestone et al., 2019a, 2019b).

about more explicit rules and requirements that should be put in place to guide AI research and development.

However, the translation of high-level principles into ground-level rules and requirements is still a time-consuming and cognitively demanding task. It requires the careful consideration of a principle's definition, the identification of relevant scenarios, engagement with the way ethical issues appear in practice, the articulation of specific guidelines (rules, checklists, etc.) for each scenario, and plans for implementation and enforcement. Decisions at each of these steps are also inherently controversial which can generate reluctance to attempt further specification.

Brent Mittelstadt (2019) explains that the translation of biomedical principles into medical practice has enjoyed some success because the medical profession benefits from a unified goal (to treat disease and alleviate patient suffering) in addition to a well-defined range of activities and established systems to facilitate in the deliberation and articulation of explicit rules and requirements (e.g. internal review boards). In contrast, Mittelstadt argues, the goals of AI developers vary widely depending on the context toward which the technology will be applied (e.g. medicine, criminal sentencing, finance, military applications, etc.), and the field currently lacks infrastructure to support in the translation and implementation of AI ethics principles. Mittelstadt doubts principles can play an important role in AI ethics governance and warns that the risk of "ethics washing" by institutions claiming to adopt AI ethics principles is high. Therefore, to ensure the development of safe and beneficial AI, AI governance should rely primarily on explicit policy requirements such as auditing and record-keeping requirements.

I agree that ethics washing is a serious concern, especially if self-adherence to ethics principles alone is expected to underpin AI governance. Indeed, even in the medical context there is some pushback against principlism. For example, following on Beauchamp and Childress's articulation of the four biomedical ethics principles, Clouser and Gert (1990) led a cautionary opposition arguing that "at best, 'principles' operate primarily as checklists naming issues worth remembering when considering a biomedical moral issue. At worst 'principles' obscure and confuse moral reasoning by their failure to be guidelines and by their eclectic and unsystematic use of moral theory" (220).[2] However, arguments such as Mittelstadt's and Clouser & Gert's that reference shortcomings in the "start-point" function of principles tend to overlook a second key function of high-level principles.

2.2 Principles Underpin Professional Culture

A less discussed, though no less important role of AI ethics principles is in underpinning cultural norms and values. Call this the "cultural influence" function of principles.

Consider again the biomedical principle of nonmaleficence. Nonmaleficence is not repeated to medical students because it is particularly profound or because

[2] Also see Davis (1995).

it outlines some useful instructions. Rather, without being translated into explicit rules and requirements, high-level principles like nonmaleficence, beneficence, and respect for patient autonomy define a mindset that influences how practitioners construe the challenges they face and the solutions they entertain.

For instance, a recent white paper released by the Word Economic Forum (2020) identifies a key function of high-level AI principles as providing a common vocabulary with which AI developers discuss design challenges and contemplate potential impacts and risks. In turn, that common vocabulary influences the kinds of solutions considered. For example, plausibly, AI developers who think primarily in terms of optimization and efficiency will take a different angle to conceptualizing a medical diagnostic system than developers who have inclusivity and explainability on their minds. Viewed as a tool for framing practitioner mindsets, it does not so much matter how many high-level AI principles there are or how exactly they are delineated or defined. What matters is that they are consistently engaged with and widely discussed and debated.

Yet more strongly still, principles can also be used to establish behavioral norms within a professional community. Norms influence individual behavior through social pressure and internalized values.[3] For instance, nonmaleficence may seem like an obviously good principle for medical professionals to abide by, yet Beauchamp & Childress's, 1979 articulation of the classic biomedical principles — nonmaleficence, beneficence, autonomy, and justice — was prompted by a series by clear breaches such as the Tuskegee syphilis trials (Beecher, 1966). Following its rearticulation as one of four central medical tenets, nonmaleficence was reinforced as a defining feature of the medical profession. For a medical professional to disregard it would be to invite social ridicule and risk rejection from the professional community.

Other principles are less obvious to begin with. For instance, prior to the articulation of the modern biomedical principles, the mantra "doctor knows best" more accurately described the medical psyche than respect for patient autonomy and informed consent. The act of introducing a new principle, broad through it may be, can help initiate a cultural shift within the professional community. This phenomenon may also occur in the context of AI ethics. The fast-moving Silicon Valley culture in which AI research and development is largely embedded primarily values efficiency, optimization, and scale (Thompson, 2019: 21–23). Many proffered AI principles such as fairness, accountability, explainability, inclusivity, and transparency challenge this status quo and could be used to instigate a similar cultural shift toward the prioritization of safety, responsibility, and human beneficence.

[3] The formation and enforcement of social norms in a social setting is a complex process. I direct interested readers to Coleman (1994: chapters 10 and 11) for a more thorough discussion.

3 Why Care About Culture?

The importance of principles as tools for influencing AI developer culture should not be downplayed. Establishing a desirable culture for AI research and development is of central importance to effective AI governance for two reasons.

First, cultural norms and values go beyond explicit rules and requirements. If rules set out the letter of the law, then norms and values are the spirit. Not every decision a physician or AI developer might face can be prescribed, and where formal rules do not dictate specific action, professional values and norms of conduct can provide overarching guidance to fill in the gaps. This is illustrated, for example, by the use of general values to guide medical decision-making in pre-clinical emergency settings where time and access to more explicit guidance is limited (Torabi et al., 2018).

Second, the alignment of cultural norms with policy goals is key to the uptake and efficacy of explicit rules and regulations. Individuals are more likely to adhere to extrinsic measures they believe in. As Seth Baum (2017) explains, explicit rules and requirements will often generate extra work, and time-consuming requirements like checklists, self-evaluations, and impact statements can foster resentment among practitioners if they are seen as unnecessarily restricting or superfluous. This resentment may, in turn, lead to rejection of the very principles being promoted (Pettit, 2002).

On the other hand, extrinsic rules and regulations have greater efficacy if they appeal to the cultural norms and values held by the communities to which they are being applied. Cultural norms motivate individual behavior through external social pressures and, if internalized, through internally generated rewards and punishment (e.g. shame or pride) (Coleman, 1994). If a person's internalized norms align with policy goals, then she will be more motivated to respond to the letter and spirit of the policy. For instance, consider the requirement that physicians obtain informed consent from patients before performing any procedure. A physician who has internalized the biomedical norm of "respect for patient autonomy" will be more motivated to ensure her patients understand their treatment options and have had all their concerns addressed, while the physician who harbors a "doctor knows best" mentality will more likely aim only to satisfy the minimum requirements for informed consent, to do so begrudgingly, and to cut corners where possible. In the case of AI, we may similarly expect that any rules and requirements will be more effective and more willingly adhered to by AI researchers and developers if those extrinsic measures align with cultural norms internalized by the community.

4 Conclusion: a Dual Approach to AI Governance

The utility of principles in their "start-point' function is limited due to the difficulty of translating high-level principles into ground-level practices. However, principles still have a key role to play for AI ethics governance in their "cultural

influence" function as catalysts for building cultures of responsibility and beneficence among AI developers. Of course, cultural change is a complicated process. Encouraging AI developers to internalize new values and to alter their social norms to align with policy goals is not merely a matter of posting a set of principles on the wall. Rather extrinsic rules and requirements play an important role in reinforcing values and 'nudging' norms to evolve. Inversely, internalized norms and values influence the efficacy of extrinsic measures, and the involvement of individual practitioners in the conceptualization, communication, and enforcement of extrinsic measures will be key to facilitating their implementation.

Principles alone should not be expected to govern AI, but nor should rules and requirements. Explicit regulation is necessary to prevent ethics washing and to ensure minimum ethical standards are met, but alignment with cultural norms is key to ensuring that explicit rules and requirements achieve their full potential. If the aim of AI governance is to ensure the responsible research, development, and deployment of safe and beneficial AI, then I posit that the most effective AI governance strategy is a dual approach; explicit rules and regulations should be buttressed by principle-based initiatives encouraging cultural change within AI research and development communities to deemphasize the cultural norms like optimization and scale that can be antagonistic to the goals of AI governance, and to promote values like beneficence and responsibility.

Acknowledgements The author would like to thank Stephen John, Angeliki Kerasidou, and two anonymous reviewers for their helpful comments on earlier drafts of this publication.

Author contribution ES is responsible for the full content of this publication.

Funding The author's research is made possible by financial support from the Cambridge Trust, The University of Cambridge Department of History and Philosophy of Science, and Trinity Hall college, Cambridge.

Availability of data and material Not applicable. No datasets were generated during and/or analysed in the preparation of this publication.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication The author consents to publication and the terms and conditions of the editors.

Competing interests The author declares no competing interests.

References

Baum, S. D. (2017). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society, 32*, 543–551. https://doi.org/10.1007/s00146-016-0677-0

Beauchamp, T., & Childress, J. (1979). *Principles of biomedical ethics*. Oxford University Press.

Beecher, H. K. (1966). Ethics and clinical research. *New England Journal of Medicine, 274*(24), 1354–1360.

Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review, 51*, 399–436.

Cave, S., Whittlestone, J., O hEigheartaigh, S. & Calvo, R. A. (2021). *Using AI ethically to tackle covid-19. BMJ, 372*(364). https://doi.org/10.1136/bmj.n364

Clouser, K. D., & Gert, B. (1990). A critique of principlism. *The Journal of Medicine and Philosophy, 15*, 219–236.

Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

Davis, R. B. (1995). The principlism debate: A critical overview. *The Journal of Medicine and Philosophy, 20*, 85–105.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., . . . Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A, 376*, 20180089.

Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI - Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society, 1*(1), 34–47.

Pettit, P. (2002). Instituting a research ethic: Chilling and cautionary tales. In P. Pettit (Ed.), *Rules, Reasons, and Norms*. Oxford Scholarship Online: Oxford University Press.

Shibata, B. (2017). An ethical analysis of euthanasia and physician-assisted suicide: Rejecting euthanasia and accepting physician assisted suicide with palliative care. *Journal of Legal Medicine, 37*(1–2), 155–166. https://doi.org/10.1080/01947648.2017.1303354

Thompson, C. (2019). *Coders: Who they are, what they think, and how they are changing our world*. Pan Macmillan Press.

Torabi, M., Borhani, F., Abbaszadeh, A., & Atashzadeh-Shoorideh, F. (2018). Experiences of pre-hospital emergency medical personnel in ethical decision-making: A qualitative study. *BMC Medical Ethics, 19*(1), 95. https://doi.org/10.1186/s12910-018-0334-x

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019a). The role and limits of principles in AI ethics: Towards a focus on tensions. *Proceedings of the 2019a AAI/ACM Conference on AI, Ethics, and Society.* https://doi.org/10.1145/3306618.3314289

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. & Cave, S. (2019b). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. Nuffield Foundation.

World Economic Forum (2020). *Ethics by design: An organizational approach to responsible use of technology.* Retrieved from: http://www3.weforum.org/docs/WEF_Ethics_by_Design_2020.pdf. Accessed 25 March 2021