DOI:10.1145/3529751

Henrik Skaug Sætra, Mark Coeckelbergh, and John Danaher

Viewpoint
The AI Ethicist's Dirty Hands Problem

Attempting to balance sometimes-conflicting interests.

IMAGE BY ANDRIJ BORYS ASSOCIATES, USING SHUTTERSTOCK

Assume an AI ethicist uncovers objectionable effects related to the increased usage of AI. What should they do about it? One option is to seek alliances with Big Tech in order to "borrow" their power to change things for the better. Another option is to seek opportunities for change that actively avoids reliance on Big Tech.

The choice between these two strategies gives rise to an ethical dilemma. For example, if the ethicist's research emphasized the grave and unfortunate consequences of Twitter and Facebook, should they promote this research by building communities on said networks? Should they take funding from Big Tech to promote the reform of Big Tech? Should they seek opportunities at Google or OpenAI if they are deeply concerned about the negative implications of large-scale language models?

The AI ethicist's dilemma emerges when an ethicist must consider how their success in communicating an identified challenge is associated with a high risk of decreasing the chances of successfully addressing the challenge. This dilemma occurs in situations in which the means to achieve one's goals are *seemingly* best achieved by supporting that which one wishes to correct and/or practicing the opposite of that which one preaches.

This dilemma is not just some fanciful thought experiment, however. For example, the ethicist Timnit Gebru tried chipping away at Google from the inside, but after authoring an article on the dangers of large-scale language models,[1] she was seemingly forced out of her job.[5] She has now founded a new *independent* research institute, and has expressed the need for AI ethics to be divorced from Big Tech.[3] However, independent institutes and academia often receive funding from big tech to support their project.[4] The ACM Code of Professional Ethics[a] urges practitioners to reflect on the impacts of their work and to support—consistently—the public good. The question, then, is *How*?

The Ethicist's Strategies

Consider the avenues of action available to a concerned AI ethicist: they can work from within the system or they can work from outside. The *system*, here, refers not just to individual AI companies, but to the larger ecosystem of companies, social structures, and political arrangements that generate the negative impacts in question. Surveillance capitalism[11] is a prime example of such a system, based on collecting and actioning personal data. It is enabled not just by individual companies but by the economic, regulatory, political, and social system. Furthermore, this sys-

[a] See https://www.acm.org/code-of-ethics

tem is sustained by ideological forces that reify and reproduce it, making structural reforms challenging. This problem has long been noted by critical media theorists.

Working within the system could entail working for the tech companies that are integral to the system, accepting positions, funding, and support from these companies and/or actively using the technologies on which the system is built. Effecting change from the outside entails abstaining from using these technologies, and choosing to conduct research and seek change either independently or by working for actors who are not part of or closely linked to the system. What are the benefits and costs of these strategies?

Strategy 1: Changing the system from within. One benefit of Strategy 1 is it provides a way to be close to relevant sources of power and to guide and change the exertion of such power in a beneficial way. When successful, this can be both an effective and nonconflictual path toward change, as the closeness to those with the power to make change happen makes direct dialogue and persuasion possible. Furthermore, the AI ethicist could, in theory, seek to become a "parasite" who accesses the system in order to change it, while actively avoiding strengthening it or giving anything back.[10] From this, one could also consider more insidious forms of subversive engagement from within, including more active forms of sabotage.

One justification for this strategy is that it leverages existing sources of power. Access to power of some sort is necessary for effecting change, and power is not necessarily bad or limiting. It can be enabling and deeply social. An individual's power always depends on the power of others, which may limit us but may also enable us. The ethicist's success depends on how their power interacts with the power of others. In the first strategy, the ethicist seeks to ally with powerful actors within the system in order to reach their goals.

In Sætra, Coeckelbergh, and Danaher[9] we use three ethical theories—virtue ethics, consequentialism, and deontology—to highlight what the ethicist allying with Big Tech must keep in mind. Virtue ethics guides us toward considerations about moral character, and whether or not certain actions are in accordance with acknowledged virtues. It seems particularly important to focus on practicing that which one preaches, which is central to the ethicist who sees grave dangers associated with social media, while wanting to use social media to warn against these dangers. While not practicing what one preaches might be morally wrong, it might also entail a loss of *ethos*—legitimacy and credibility—that can hurt the ethicist's effectiveness in conveying their message.

Consequentialism guides us toward the good or bad outcomes that might be associated with our actions. This might be thought to allow the ethicist to use the power of Big Tech to warn against ethical dangers, on the grounds that doing so serves a greater good. However, by allying with Big Tech and using their technologies, the ethicist risks: supporting the companies and making their products more attractive; supporting the platforms they wish to undermine by driving traffic, ad sales, and so forth; and legitimizing and becoming complicit with the platforms they criticize. In this respect, ethicists should be particularly wary of "ethics-washing," as Big Tech uses them to demonstrate that these companies are open to criticism and that they take ethics seriously. The net effect of this is potentially to deepen the problems and block other means of achieving change.

Finally, deontology guides us toward our duties and toward the consideration that the end does not necessarily justify the means. Deontology, at least one popular form, entails seeking universal rules for action. But making it a general rule that an ethicist ought to ally with Big Tech could be self-undermining. If everyone allies with Big Tech, the actors involved grow more powerful, their platforms gain legitimacy, and the ethical challenges are not effectively met.

Consider the avenues of action available to a concerned AI ethicist: they can work from within the system or they can work from outside.

Strategy 2: Circumvent the old to make way for the new. Working from outside the system can seem daunting but may lead to more lasting change. As Audre Lorde stated, "The master's tools will never dismantle the master's house."[6] Systems shape perceptions of what is possible and desirable, and any change effected from within will partly be a result of the logic of the system in question.[7] Working from within the system, ethicists will arguably be prone to chase relatively minor problems and become preoccupied with the technical details of existing solutions. They will become beholden to the system's ideology. This might make it difficult to perceive and understand the real problems.

However, rejecting any sort of alliance with Big Tech entails downsides of its own. Independence from power might be important, but it can entail insignificance. Working outside the system can minimize the ethicist's voice, audience, and potential for impact. Chipping away from outside the system can be a lonely and frustrating task. One may not be taken seriously by those with the power to change the system. One may be perceived as a crank or nuisance. Furthermore, it may be practically impossible to achieve true independence from the system. One of the main concerns with Big Tech is how endemic it is to our economic, social, and political lives—including our academic lives and democratic systems. In short, the ethicist who disconnects from any sort of power runs the risk of *marginalization*, *impotence*, or *ineffectiveness*.

Nevertheless, there are good reasons for choosing to work from the outside, and the ethical theories we rely on highlight some of these. From a virtue ethics perspective, ethicists working outside the system avoid complicity with the systems they find problematic, and enhance their ethos. From a consequentialist perspective, they avoid strengthening the system and being used as a front for ethics washing. In particular, they avoid becoming entangled in an incentive structure that makes efforts to achieve radical change both risky and often self-defeating. While being employed by, taking funding from, or relying on the tools of Big Tech does not invalidate one's research, it seems likely that solutions proposed by those in such positions are more focused on changing and adjusting the current system rather than overthrowing it. From a deontological perspective, AI ethicists are better able to do their duties as critics of the system and avoid being seen as using the tools of that which they seek to change. Not being part of the system, their moral conscience is clear.

The Need for More than AI Ethics

Our analysis of the ethicist's dilemma shows why close ties with Big Tech can be detrimental for the ethicist seeking remedies for AI related problems.[9] It is important for ethicists, and computer scientists in general, to be aware of their links to the sources of ethical challenges related to AI. One useful exercise would be to carefully examine what could happen if they attempted to challenge the actors with whom they are aligned. Such actions could include attempts to report unfortunate implications of the company's activities internally, but also publicly, as Gebru did. Would such actions be met with active resistance, with inaction, or even straightforward sanctions? Such an exercise will reveal whether or not the ethicist *feels* free to openly and honestly express concerns about the technology with which they work. Such an exercise could be important, but as we have argued, these individuals are not necessarily positioned to achieve fundamental change in this system.

In response, we suggest the role of government is key to balancing the power the tech companies have through employment, funding, and their control of modern digital infrastructure. Some will rightly argue that political power is also dangerous.[2] But so are the dangers of technology and unbridled innovation, and private corporations are central sources of these dangers. We therefore argue that private power must be effectively bridled by the power of government.[8] This is not a new argument, and is in fact widely accepted. For example, government intervention for the sake of correcting market failure is normally accepted. The constructive power of government must also be embraced, and Strategy 2 and the path of stronger *political* regulation, rather than only incentives, ethics-from-within, and self-regulation is here advocated as necessary, if not sufficient, in order to solve some of the ethical problems generated by AI. Furthermore, this necessitates clearer and more transparent boundaries between policy advisors and the targets of regulation.[3]

We suggest the role of government is key to balancing the power the tech companies have through employment, funding, and their control of modern digital infrastructure.

While working from within the system may allow the ethicist to slightly change the direction of company strategies and policies, working from outside seem to allow for a larger canvas on which to sketch new and fundamentally different solutions. However, a potential drawback of this avenue is that the goal of achieving social change is not reached because the degree of interaction with the system and the rest of society is so low that no or only a weak alliance can be forged.

However, a very high degree of interaction between Big Tech and the political and legal institutions is even more problematic. The tight linkages between the system and the political domain, which we argue is crucial for controlling and correcting the system, when necessary, makes it very difficult to find space "outside" the system and work to reform it at a fundamental level. This is one of the major issues with AI ethics at the moment, when so much of the Big Tech economy and indeed so much of society is dependent on the system, that it is difficult for the AI ethicist to find sufficient support and space outside the system—even in academia. To avoid this, one must be strategic in terms of building alliances, while at the same time avoiding the disadvantages of getting too entangled in the system

More work is needed on the relationships between power, social change, and technology, including on a more relational conception of power. Power is not only dangerous; it is also necessary to change the world. We conclude that it is necessary to not only pursue ethics-from-within, as seeking change from the outside is crucial for effecting fundamental political change. Since both strategies have clear advantages and disadvantages, a combination is desirable.

References

1. Bender, E.M. et al. On the dangers of stochastic parrots: Can language models be too big. In *Proceedings of FAccT,* 2021; https://bit.ly/3gdvKTp
2. Chomanski, B. The missing ingredient in the case for regulating big tech. *Minds and Machines* (2021) 1–19; https://bit.ly/3Aou4Ou.
3. Clarke, L. et al. How Google quietly funds Europe's leading tech policy institutes. *The New Statesman* (2021).
4. Gebru, T. For truly ethical AI, its research must be independent from big tech. *The Guardian* (2021).
5. Hao, K. I started crying: Inside Timnit Gebru's last days at Google—and what happens next. *MIT Technology Review* (2020).
6. Lorde, A. *The Master's Tools Will Never Dismantle the Master's House.* Penguin, U.K., 2018.
7. Næss, A. *Ecology, Community and Lifestyle: Outline of an Ecosophy.* Cambridge University Press, 1989.
8. Sætra, H.S. *Big Data's Threat to Liberty.* Academic Press, 2021.
9. Sætra, H.S. et al. The AI ethicist's dilemma: Fighting Big Tech by supporting Big Tech. *AI and Ethics* (2021); https://bit.ly/3EIzqpU.
10. Serres, M. *The Parasite.* U of Minnesota Press, 2013.
11. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power: Barack Obama's Books of 2019.* PublicAffairs (2019).

Henrik Skaug Sætra (henrik.satra@hiof.no) is an associate professor at Østfold University College in Norway.

Mark Coeckelbergh (mark.coeckelbergh@univie.ac.at) is a professor of Philosophy at the University of Vienna in Austria.

John Danaher (john.danaher@nuigalway.ie) is a senior lecturer at the University of Galway in Ireland.