Canadian Journal of Philosophy (2025), 1–17
doi:10.1017/can.2025.9

ARTICLE

How Can We Know if You are Serious? Ethics Washing, Symbolic Ethics Offices, and the Responsible Design of AI Systems

Justin B. Biddle, John P. Nelson and Olajide E. Olugbade

School of Public Policy, Georgia Institute of Technology, Atlanta GA, USA
Corresponding author: Justin B. Biddle; Email: justin.biddle@pubpolicy.gatech.edu

Abstract
Many AI development organizations advertise that they have offices of ethics that facilitate ethical AI. However, concerns have been raised that these offices are merely symbolic and do not actually promote ethics. We address the question of how we can know whether an organization is engaging in ethics washing in this way. We articulate an account of organizational power, and we argue that ethics offices that have power are not merely symbolic. Furthermore, we develop a framework for assessing whether an organization has an empowered ethics office—and, thus, is not ethics washing via a symbolic ethics office.

Keywords: ethics and governance of artificial intelligence (AI); organizational ethics; responsible research and innovation; social epistemology; values in science and technology

1. Introduction

The past few years have witnessed an explosion of interest in, and discussion about, emerging technologies involving artificial intelligence (AI). Governments, firms, civil society organizations, and publics are experimenting with new AI technologies and debating their implications, expressing both angst and optimism. As contributions to these debates, stakeholder organizations have created and disseminated numerous AI ethics documents (e.g., Jobin, Ienca, & Vayena, 2019; Schiff et al., 2021). Produced by AI development firms, governments, and non-profit organizations, these documents identify ethical principles or values—such as beneficence, non-maleficence, respect for persons, justice, and explicability (e.g., Floridi and Cowls, 2019)—that should guide the development and use of AI.

At the same time that AI development organizations are advertising their commitments to ethics and social responsibility, however, some are acting in ways that have elicited charges of "ethics washing." One example is Google's Advanced Technology External Advisory Council (ATEAC), which was announced on March 26, 2019, and advertised as helping to implement Google's AI Principles, "an ethical charter to guide the responsible development and use of AI in our research and products" (Walker, 2019). ATEAC would "consider some of Google's most complex challenges that arise under our AI Principles, like facial recognition and fairness in machine learning, providing diverse perspectives to inform our work" (Walker, 2019). ATEAC was short-lived. After only 9 days, Google dissolved the council over external and internal critiques of its membership (Piper, 2019). For many commentators, the episode raised concerns about "ethics washing"—a term we discuss at greater length below. For example, in a December 2019 article for the *MIT Technology Review* entitled, "In 2020, Let's Stop AI Ethics-Washing and

Actually Do Something," Karen Hao referred to the ATEAC fiasco as "the most acute example" of ethics washing in AI (Hao, 2019).

The ATEAC event is not an isolated incident. Other technology firms have created offices of ethics and responsibility, touted their creation as evidence that they are behaving ethically, and then eliminated the offices; prominent researchers in AI companies who have raised alarms about bias and injustice have been fired, as have entire teams of researchers dedicated to examining the societal implications of AI (e.g., De Vynck & Oremus, 2023; Knight, 2022; Schiffer & Newton, 2023). Current and former employees of OpenAI and Google DeepMind recently published an open letter, entitled "A Right to Warn about Advanced Artificial Intelligence," which exposed "disparagement clauses" restricting employees' ability to discuss ethical concerns. "[B]road confidentiality agreements block us from voicing our concerns, except to the very companies that may be failing to address these issues" ("A Right to Warn", 2024). Incidents such as these, in conjunction with many highly-publicized examples of biased or harmful AI systems, have led many to argue that contemporary AI ethics discourse in its entirety amounts to little more than ethics washing (e.g., Munn, 2022).[1]

Why, more specifically, do ATEAC and other similar episodes raise concerns about ethics washing? While many aspects might be highlighted, one noteworthy feature pertains to organizational structure, including the structure of ethics offices such as ATEAC and their relations to research and development (R&D) activities. ATEAC was external to Google; it consisted of unpaid volunteers; it met infrequently (4 times a year), and it provided non-binding recommendations that Google's R&D teams were free to ignore (Piper, 2019). The office was supposed to facilitate ethical AI, but *it lacked power*—not because of any personal characteristics of the individuals within the office, but because of the structure of the organization and the place of the office within that structure. Given that other AI development organizations have created offices of ethics that purportedly work to align AI systems with ethical values, it is important to inquire into these offices and their relations to R&D activities.

The high-level questions that motivate this article include how we can know whether an AI organization that signals a commitment to ethics is engaging in ethics washing and, more generally, what it means for an AI organization to act responsibly in its R&D practices. These questions are important from many perspectives, including the practical epistemology of science/technology communication, the ethics of emerging technologies, and also from broader societal perspectives. However, these questions are very broad, and we cannot hope to treat them rigorously within a single essay. In this article, we will focus on a more specific question: *If an AI organization has an office of ethics that is advertised as facilitating ethical AI, how can we assess whether that office is merely symbolic or performative?*

By a "symbolic" or "performative" ethics office, we mean an office that advertises itself as promoting ethical values but does not actually do this. It might act as a marketing instrument to prevent or mitigate reputational harm or forestall government attempts to regulate the industry, or generate positive press about the organization. It might even consist of well-intentioned individuals who make thoughtful recommendations to other organizational units. However, it does not affect organizational change by promoting goals of ethics or social responsibility (c.f., Meyer & Rowan, 1977; Edelman, 1992; Westphal, Gulati, & Shortell, 1997). We address the question of how we can know whether an organization's ethics office is merely symbolic because we believe that this is a common way that technology organizations, including AI development organizations, engage in ethics washing.

The structure of this article is as follows. We begin, in Section 2, by articulating an account of ethics washing and organizational power. We argue that an ethics office is not merely symbolic if and only if it meets the criteria of power specified in this account. In our view, the ATEAC case is a good illustration of the strategy of ethics washing via symbolic ethics offices, and we return to it in

[1] There is a large literature on bias and values in AI and machine learning. Discussions in philosophy include Biddle (2022, 2023), Fazelpour and Danks (2021), and Johnson (2021).

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

Section 2.[2] In Section 3, we develop an assessment framework that can be used to make qualified assessments about whether an organization has an ethics office that has power—and, thus, is not merely symbolic. We address the scope and limitations of our argument in Section 4, and we conclude in Section 5.

2. Ethics Washing and Organizational Power

According to standard usage, ethics washing refers to the practice of signaling a commitment to ethics without genuinely having such a commitment or sufficiently putting it into practice. The concept of ethics washing is analogous to greenwashing, which refers to the practice of signaling a commitment to sustainability and the environment, without sufficiently having such a commitment or putting it into practice. At present, there is a relatively scant philosophical or scholarly treatment of ethics washing. As of April 2024, the academic database Web of Science indexed only 20 English-language documents containing the term and grammatical variants in their titles, abstracts, or keywords. Of these English-language documents, 15 include definitions of ethics washing (Ahmad et al., 2021; Aitken et al., 2021; Ali, Christin, Smart, & Katila, 2023; Bietti, 2020; Bürger, Amann, Bui, Fehr, & Madai, 2024; Kotliar & Carmi, 2023; Lohne, 2021; McMillan & Brown, 2019; Papyshev & Yarime, 2022; Saltelli, Dankel, Di Fiore, Holland, & Pigeon, 2022; Siapka, 2022; Steinhoff, 2024; Vică, Voinea, & Uszkai, 2021; Wilson et al., 2024; Wright, 2023). Some of these documents refer to others (not listed in the Web of Science database) that contain definitions of ethics washing (Floridi, 2019; Johnson, 2021; Metzinger, 2019). Explicitly and implicitly, these articles use several overlapping notions of ethics washing. However, the usage in each is consistent with the following definitional schema: ethics washing is *signaling about ethics plus X, where X is either an intention or behavior that is not sufficiently aligned with one's signaling.*

To engage in ethics washing, an organization must signal a commitment to ethics. Signaling activities might include marketing campaigns that advertise ethical commitments, recruiting "ethics experts" as consultants and publicizing their activities, and the like. Given that ethics signaling is necessary for ethics washing, unethical behavior need not involve ethics washing. Organizations that behave unethically but refrain from ethics signaling are not ethics washing. Such organizations might be acting badly, but if they are not advertising themselves as behaving ethically, then they are not ethics washing.

In addition to signaling about ethics, organizations that engage in ethics washing do something more—they behave, or have intentions, that are not sufficiently aligned with their signaling. Some of the problematic additions highlighted by other commentators include:

- Perpetuating falsehoods or exaggerations about ethical intent or behavior (a behavior) (Aitken et al., 2021; Ali et al., 2023; Bietti, 2020; Floridi, 2019; Johnson, 2021; Steinhoff, 2024; Vică et al., 2021)
- Failing to act in ways required by one's stated ethical commitments (a behavior) (Ahmad et al., 2021; Aitken et al., 2021; Bietti, 2020; Bürger et al., 2024; Floridi, 2019; Metzinger, 2019; Papyshev & Yarime, 2022; Saltelli et al., 2022; Siapka, 2022; Vică et al., 2021; Wright, 2023)
- Ethics signaling for purposes of public relations (an intention) (Ahmad et al., 2021; Aitken et al., 2021; Bietti, 2020; Floridi, 2019; Papyshev & Yarime, 2022; Steinhoff, 2024; Wilson et al., 2024)
- Ethics signaling for purposes of forestalling regulation (an intention) (Kotliar & Carmi, 2023; Lohne, 2021; McMillan & Brown, 2019; Metzinger, 2019; Papyshev & Yarime, 2022; Saltelli et al., 2022; Siapka, 2022; Vică et al., 2021)

[2] There are other examples of ethics offices that facilitate ethics washing, which are more complicated in many respects than ATEAC. We plan to discuss these in future work.

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

Amongst the discussions of ethics washing that we have reviewed, we find behavior-based accounts, intention-based accounts, and mixed accounts that involve both behaviors and intentions.

In this article, we adopt a behavioral account of ethics washing, as *signaling a commitment to ethics, without acting in ways that are sufficiently aligned with that signaling.*[3] We do this because of the high-level epistemic goals that motivate our project, including (again) understanding how we can know whether an AI organization that signals a commitment to ethics is engaging in ethics washing and, more generally, how we can know whether an AI organization is acting responsibly in its R&D practices. To make such assessments, we must make reference to behavior by organizations that can be observed.

In adopting a behavioral account, we are not arguing that intention-talk is illegitimate. We are not arguing that accounts of ethics washing that reference intentions are necessarily problematic. However, our project is assessment-driven. If we are to assess whether an organization is ethics washing, our assessment should be grounded in behaviors that can be observed. Hence, our adoption of a behavioral account.

There are interesting conceptual questions related to ethics washing, many of which are beyond the scope of this article. We will highlight three of them here, as doing so will help to clarify the scope of our argument. First, our account of ethics washing does not presuppose the truth of any particular ethical theory or approach. Whether an organization is engaging in ethics washing depends upon the alignment (or lack thereof) between its signaled ethical commitments and behavior; it does not depend on the truth or acceptability of the signaled ethical frameworks. Second, given an account of ethics washing as signaling a commitment to ethics without acting in ways that are sufficiently aligned with that signaling, there are interesting questions about what counts as *sufficient* alignment. These questions fall outside of the scope of this article, as our main goal is to examine how to distinguish merely symbolic from substantive ethics offices.[4] Third, there are interesting and important questions about how ethics washing relates to neighboring moral concepts, including moral praiseworthiness and blameworthiness. These questions, again, fall outside of the scope of our argument. We now turn to the relationship between ethics offices, organizational power, and ethics washing.

2.a. What is an ethics office?

As we have discussed, an important strategy according to which AI organizations might engage in ethics washing is through the creation of ethics offices that are merely symbolic. In this section, we articulate an account of organizational power, and we argue that ethics offices that have power are not merely symbolic. Thus, if an AI development organization has an ethics office with power, it is not engaging in ethics washing in this way. Conversely, if it has an ethics office that is not

[3] This broad conception of ethics-washing also aligns with accounts of the better-established phenomenon of greenwashing, for example, as articulated by Gu and Matisoff (forthcoming).

[4] We hypothesize that judgments about sufficient alignment should consider (at the very least) two dimensions: the degree of misalignment and the societal consequences of misalignment. A greater degree of misalignment is, *ceteris paribus*, more likely to constitute ethics washing, whereas a lesser degree might be appropriately characterized as mere exaggeration. With respect to societal consequences, there are some cases in which overstating one's trustworthiness can lead to severe harms. Mischaracterizing the stringency of one's safety testing or bias mitigation efforts can result in AI systems that cause life-changing harms, including false arrest and imprisonment, deportation, and job suspension or termination. In other cases, the consequences of misalignment might be much more mundane. Cases in which the societal consequences of misalignment are severe might, *ceteris paribus*, be appropriately characterized as ethics washing. Assessing sufficiency in light of the consequences of misalignment can be seen as an extension of the now voluminous literature on inductive and epistemic risk (e.g., Biddle, 2022; Biddle & Kukla, 2017; Douglas, 2000; Havstad, 2022; Rudner, 1953). In any case, the focus of this paper is, again, on assessing whether an ethics office is merely symbolic; questions about what counts as sufficient alignment between ethics signaling and behavior fall outside of the scope of this argument. We are grateful to an anonymous referee for comments on this issue.

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

empowered, then the office is merely symbolic, which raises concerns about whether the organization is engaging in ethics washing. Before we proceed further, we must clarify what we mean by an ethics office.

Many technology firms have (or have had) offices that identify ethical values and that purportedly work to align the organization's operations with those values. The titles of such offices vary; they include Microsoft's Office of Responsible AI (Microsoft, 2022), Salesforce's Office of Ethical and Humane Use (Spiegel, 2023), and X's (formerly Twitter's) now disbanded Trust and Safety Council (Associated Press, 2022). We use the term "ethics" broadly, to include normative considerations about what counts as socially responsible behavior; what kinds of actions or characters are worthy of trust; which distributions are fair; which benefits and harms should be prioritized; which risks are acceptable and what counts as "safe enough," and the like. By an ethics office, we mean a subunit of an organization that interacts with other subunits to facilitate the alignment of organizational operations with ethical values. In the case of AI development organizations, we define an AI ethics office as *an organizational subunit that interacts with other subunits, including subunits performing research and development, to facilitate the development, implementation, or use of AI in accordance with ethical values.*

To facilitate this, an AI ethics office might, for example, identify clearly and transparently the ethical values that the organization aims to promote; articulate specific ways in which these values should be operationalized in research and development practices, including how the ethical values should affect organizational processes and decision-making; define the processes that will be undertaken in case of allegations of ethics violations, such as when employees believe that the organization is pursuing a line of research that conflicts with its ethical principles; and clarify how any such allegations will be resolved, including potential sanctions.

These are steps that an AI ethics office could take. We do not, however, build them into our definition of an ethics office, as we want to avoid defining an ethics office as one that is effective in promoting ethical values. As the example of Google's ATEAC suggests, some ethics offices can be ineffective. They might include individuals who are serious about ethics and who deliberate carefully and responsibly about ethics in AI—and they might do this even if the office itself is ineffective. An ethics office might fail to facilitate its ethics goals because other subunits are not responsive to it. An ethics office might also fail to facilitate its ethics goals because the processes that it implements to interact with other subunits are not effective at promoting the ethical aims that it espouses.

A final word is merited on the potential overlap between ethics offices and legal compliance offices. Many organizations have compliance offices to ensure that organizational behavior comports with legal requirements, and these offices engage in many similar behaviors to those of idealized ethics offices (e.g., identifying important norms to follow, issuing guidance or requirements for how to do so, assessing compliance with those requirements). To the extent that ethical values and legal requirements overlap, the actions of an ethics office and a legal compliance office could overlap. Given this overlap, we can imagine an AI organization that has an office of ethics and compliance, which advertises that it follows ethical values to the extent that is required by law. If we were to assess this organization for ethics washing, we would want to know whether its behaviors are aligned with its signaling—that is, whether it is behaviors are in compliance with the law. However, many organizations—including many AI organizations—advertise that they are doing more than this. They advertise that they are promoting values of social responsibility, equity, transparency, privacy protection, and others—and doing so in ways that go beyond what the law requires (e.g., Schiff, Borenstein, Biddle, & Laas, 2021). We do not build into our definition of an ethics office that it promotes values that go beyond legal compliance. But AI organizations that tout their commitments to ethics are typically asserting or implying that they are doing more than merely complying with the law—and if their actions do not sufficiently align with their signaling, then they are engaging in ethics washing.

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

2.b. What is organizational power?

To develop the argument that ethics offices with power are not merely symbolic, we must address the question of what constitutes power within an organization. For the purposes of this discussion, we characterize power in terms of the ability to influence another to do something they would not otherwise do (Dahl, 1957). Because we are interested in explicating what is required for an ethics office to have power, we do not focus primarily on accounts that locate power in individual persons. Rather, we examine the structural conditions under which organizational units have power. The theory that we draw upon is the strategic contingency theory, as elaborated by Hickson, Hinings, Lee, Schneck, and Pennings (1971), which is an influential account of intraorganizational power in organization theory. Strategic contingency theory holds that intra-organizational units have power to the extent that they control strategic contingencies, where a contingency is defined as "a requirement of the activities of one subunit which is affected by the activities of another subunit" (ibid., 222). On this account, an organizational subunit has power if and only if it has some control over decisions or resources that impact other subunits, such that its elimination would result in significant disruption to the overall organization. More specifically, the following conditions are necessary and jointly sufficient for an intra-organizational unit to have power: (1) effectively coping with uncertainties in their organizational environment, (2) not easily substitutable, and (3) centrality. If an organizational unit possesses each of these to a nonzero degree, then it has power; how much power a unit has depends on the degree to which it possesses each characteristic.

Consider, first, how an AI ethics office might effectively cope with uncertainty. There is significant uncertainty involved in developing AI systems that satisfy ethical principles. Depending on which systems are developed, and how they are designed, deployed, and regulated, AI systems will have uncertain impacts on the well-being of various communities; opportunities for democratic deliberation; education; surveillance and privacy risks; labor markets, career prospects, and job satisfaction, and many other aspects of life. There is significant uncertainty in predicting how design decisions in early-stage R&D activities will impact users and communities downstream; there is significant uncertainty in how users will respond to AI systems put on the market—including whether consumers will use the systems in ways specified by developers; there is uncertainty in how regulators and other stakeholders will perceive the impacts of AI systems and the organizations that produce them. Decisions about development and deployment have uncertain downstream impacts on society, which in turn impact the future prospects of the organization of which the ethics office is a part. Thus, if the ethics office influences decisions about development and deployment, it is coping with uncertainty—it is managing uncertain decisions, the outcomes of which have significant impacts on the organization as a whole.

Second, the office is not easily substitutable. If an office has expertise that is not otherwise easily accessible to the organization, then the office is not easily substitutable. In the case of an empowered AI ethics office, the members of the office would likely have relevant ethics expertise, as well as sufficient expertise in AI to be able to identify technical decisions that involve epistemic risk and that have the potential to impact communities in morally significant ways. Furthermore, they also need to be able to communicate and work effectively with R&D personnel. Individuals trained in ethics, science and technology studies, and related fields—which might prioritize responsible governance and careful deliberation about ethically complex decisions—do not always mesh well with individuals trained in technology development fields that prioritize speed, including being first to publish and/or first to market. Thus, members of an empowered AI ethics office would not only have the relevant ethics expertise but also the social skills and interdisciplinary capabilities to work effectively with scientists, engineers, and others. If such expertise does not exist elsewhere within the organization or potential contractors, the office is not easily substitutable.

Third, an empowered ethics office is central. While the concept of centrality can be spelled out in many ways, Hickson et al. highlight two aspects that are crucial. The first is workflow pervasiveness —or the "degree to which the workflows of a subunit connect with the workflows of other subunits"

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

(ibid., 221). The second concerns the significance of the activities of the unit for the organization; "the activities of a subunit are central if they are essential in the sense that their cessation would quickly and substantially impede the primary workflow of the organization" (ibid., 222). For an ethics office to have power, it must be connected to other subunits and be viewed as indispensable to the success and proper functioning of the overall organization.

There are many ways of instantiating a "central" ethics office, in the sense of workflow pervasiveness. For example, an ethics office might be independent of R&D and act as a kind of "watchdog" to ensure that AI systems are developed and deployed ethically. Alternatively, an ethics office could be integrated into R&D by embedding ethicists in technology development teams, in ways similar to socio-technical integration research (e.g., Fisher, 2019). Or an ethics office could oversee a mandated ethics training of employees, for the purpose of distributing ethics expertise throughout the organization and fostering a culture of ethics. There are many potential ways in which an ethics office could be central.

Many real-world ethics offices, however, do not meet this centrality condition.[5] In some cases, offices that promote ethical aims—such as equity, diversity, inclusion, and responsibility—are unfortunately organizationally peripheral. In these cases, the elimination of these offices would have little impact on other organizational units or processes and the behavior of the organization as a whole. Arguably, Google's ATEAC was organizationally peripheral in this sense. In addition to the fact that it did not appear to be coping with uncertainty, it also did not appear to be organizationally central. Even if it had operated for a longer period of time, the elimination of the office would arguably not have disrupted the activities of the overall organization, including R&D, in any significant way. Given that many ethics offices do not, in practice, have power, under what conditions might they?

2.c. Under what conditions would an office of ethics have power?

For an office to control strategic contingencies, it would need to perform activities or services that are crucial to the overall functioning of the organization, such that the closure or removal of the office would significantly interfere with the ability of the organization to achieve its aims. In other words, it would need to be the case that the office is responsive to some ethics-related internal or external pressure that constitutes a strategic contingency with which an ethics body is especially positioned to cope. Under what conditions would an ethics office play such a crucial role? We envision at least three.

First, an office of ethics might control strategic contingencies given sufficient popular or market pressure. Consumers, companies, public interest groups, and others could pressure AI development companies to act more ethically. This might be done through purchasing decisions, protests, political action campaigns to regulate AI, and the like. There might also be market-driven efforts to certify that AI systems have undergone heightened ethical scrutiny and to encourage consumers to purchase systems that have been thus certified. Under such pressure, an ethics office could help an organization to avoid unethical behavior and to signal such avoidance externally.

Second, governments could pressure AI companies to adhere to ethical standards. This might be done through regulation, incentives, or other means. For example, the European Union has passed the Artificial Intelligence Act, which has as one of its aims to ensure "a high level of protection of health, safety, fundamental rights enshrined in the Charter of Fundamental Rights, including democracy, the rule of law and environmental protection, against the harmful effects of artificial intelligence systems" (Chapter 1, Article 1). This Act would ban AI systems that pose "unacceptable

[5] As noted earlier, many technology companies have created offices of ethics or responsibility, advertised their creation as evidence that they are taking ethics seriously, and subsequently eliminated the offices—apparently without significantly disrupting the operations of the organization as a whole (e.g., De Vynck & Oremus, 2023). These cases constitute evidence that the ethics offices do not meet the centrality condition.

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

risk" to fundamental rights, and it would impose pre-market "conformity assessments," among other requirements, on systems deemed to be "high risk." In a regulatory environment such as this, an ethics body could help to ensure organizational compliance and mitigate the risk of reputational harm, financial penalties, and other sanctions.

In the third circumstance, powerful persons within or important to an organization—for example, top management, major investors or funders, or difficult-to-replace workers—could demand that ethical standards be met and threaten to divest, pull funding, strike, quit, or otherwise disrupt organizational operations if the organization failed to do so. In this case, an ethics office could help to prevent such disruptions by facilitating organizational behavior that aligned with ethical values.

As should be clear from this discussion, it is a significant challenge to sustain these conditions in a way that would make an ethics office crucial to the overall functioning of an organization—especially of a private, for-profit firm. It is difficult for public interest groups to maintain significant pressure on technology companies, particularly if they have limited and/or uncertain resources. Consumers and the publics also face difficulties in accessing accurate information about what AI firms are doing, and the respects in which their actions might (or might not) satisfy ethical values. Consumers and the publics are typically unable to access information to the degree that governmental regulatory agencies can. Regulatory bodies face their own challenges, however, particularly in environments where political leaders are dependent on technology firms for campaign financing. Finally, although leaders within technology firms might have good intentions about developing AI systems responsibly, they also have other interests (financial, reputational, etc.) that could constitute conflicts of interest and nudge them to cut corners with respect to ethics. In short, creating and sustaining a powerful ethics office is a major challenge.

2.d. Offices of ethics, organizational power, and ethics washing

We are now in a position to address directly the relationship between organizational power, merely symbolic ethics offices, and ethics washing. At the beginning of Section 2.a, we defined an AI ethics office as an organizational subunit that interacts with other subunits, including R&D, to facilitate the development, implementation, or use of AI in accordance with ethical values. To have organizational power is, again, to control strategic contingencies, which includes coping with uncertainty in ways that impact other organizational subunits, and doing so in ways that are organizationally indispensable and not easily substitutable. If an AI ethics office has power—if it controls strategic contingencies—then it is effective, at least to some degree, in facilitating the development, deployment, or use of AI in accordance with ethical values. Thus, it is not acting in a purely symbolic manner. Conversely, if an AI ethics office is purely symbolic, then it is not effectively acting to promote ethical AI. Empowerment is thus necessary and sufficient for concluding that an ethics office is not merely performative.

As an illustration, consider again the case of Google and ATEAC—an alleged example of ethics washing via a purely symbolic ethics office. The claim that many critics made was that ATEAC would have had no effect on decision-making processes or outcomes at Google. ATEAC, had it come into existence, would have made merely non-binding recommendations ("toothless"), and it would have done so from a position outside of the company. According to critics, the formation of ATEAC would have signaled a commitment to ethics, but when it came to decisions about the development and deployment of technologies, ATEAC would have had no influence. This could not happen if the ethics office had power. To have power, an office must be central—which is to say, again, that its workflows are connected with the workflows of other units and that its activities are "essential in the sense that their cessation would quickly and substantially impede the primary workflow of the organization" (Hickson et al., 1971, 222). ATEAC would not have been central, in this sense. Additionally, if the office had power, then it could not have been toothless. To control strategic contingencies is, among other things, to make decisions that place constraints on other organizational units. An office with power affects other organizational units, and an ethics office interacts with other subunits to facilitate the development, deployment, or use of AI in accordance with stated ethical values. An ethics office with power is one that places ethical constraints on other subunits. Thus, if an AI ethics office has power, then it is not acting in a purely symbolic manner. If it is acting purely symbolically, then it is not empowered.

One might object to our argument on the grounds that our view is too demanding. Perhaps an ethics office could do some good, even if it is not empowered in the way that we have elaborated. Perhaps the activities of the ethics office extend no further than virtue signaling—say, by posting flyers about the importance of social responsibility and equity—and this signaling influences researchers to consider impacts that they would not have otherwise considered.[6] We do not deny that this is a possibility. And we do not argue that, if an organization has an ethics office that is not empowered, it is necessarily engaging in ethics washing. We argue that if an organization has an ethics office that is not empowered, then the ethics office is acting in a merely performative manner —and this is a common pathway for ethics washing.

This conclusion is important because it provides an avenue for assessing whether an organization is ethics washing in a common way, namely via a merely symbolic ethics office. In the next section, we will provide guidance for empirically assessing whether an organization's ethics office is acting in a merely symbolic manner. As noted previously, however, there are many different ways in which an organization might engage in ethics washing—and some might involve ethics offices that are not merely symbolic. Before proceeding to the next section, we will highlight three other kinds of ethics offices that might engage in ethics washing: the *under-resourced* ethics office, the *selectively committed* ethics office, and the *co-opted* ethics office. This is not an exhaustive list; it is intended to indicate other pathways to ethics washing, as well as highlight the relevance of organizational power in this phenomenon.

First, imagine a relatively small and/or narrowly focused ethics office within a very large organization that is marketing itself as being socially responsible. The ethics office publicly espouses ethical values, it has power, and it acts in accordance with its espoused ethical values—but it does not have the capacities for monitoring and intervention that are needed in such a large organization. This office might focus on a narrow range of development activities—for example, red/yellow/green lighting early-stage research projects from an ethical point of view, or influencing late-stage decisions about whether a developed system is sufficiently safe to be disseminated or put on the market. It might focus on a narrow subset of ethical values—it might, for example, positively influence the development of AI systems that are explainable and interpretable, but it might lack the capacity to promote other ethical values. Such an ethics office would not be merely symbolic, but it would be inadequate to promote effectively the ethical values that the organization espouses. As a result, the organization could still be engaging in ethics washing.

Second, consider an ethics office that publicly espouses some ethical values, that has power, and that acts in accordance with its espoused ethical values—but only selectively, when it is sufficiently convenient to do so. The organization imposes obligations on its employees on the basis of its stated ethical principles; it institutes policies and decision-making procedures that promote the satisfaction of these ethical principles—such as policies that require ethical review of AI systems before they are disseminated publicly—and it has avenues for recourse in case these policies or procedures are violated. The ethics office copes with uncertainty, is not easily substitutable, and is central, and its adoption of ethical principles leads it, in many situations, to act in ways that better align with its principles than it otherwise would, *ceteris paribus*. However, suppose that there are regular occurrences in which the organization fails to live up to its stated ethical principles. In some situations—for example, if decision-makers are especially concerned about meeting a deadline and worry about the consequences of delay—it allows AI systems to be disseminated publicly without

[6] We are grateful to an anonymous referee for highlighting this objection.

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

undergoing ethical review (cf. Roose, 2024). The employees or teams involved in such decisions might be well-intentioned; they might believe that the AI systems that are disseminated without review are perfectly safe, that there is minimal risk of harm, and that insisting on ethical review would result in unnecessary delays. But the ethics office is, in these situations, violating its own ethical principles. Such an ethics office is not merely symbolic. It has power, and it promotes organizational adherence to ethical values in many (perhaps even most) situations. At the same time, the organization is engaging in ethics washing if it fails to act in ways that are sufficiently aligned with its ethics signaling.

An even more insidious form of ethics washing could occur if a powerful ethics office is co-opted by individuals and/or units in the organization that are acting unethically and/or illegally. Consider an ethics office that is similar, in some respects, to the selectively committed ethics office—it publicly espouses some ethical values, it has power, and it acts in ways that promote ethical values in some cases. However, the organization is also engaged in activities that are irresponsible, illegal, and/or detrimental to society, and the ethics office facilitates these activities because it is co-opted. Zaman et al. (2021) discuss the consequences of corporate misconduct of boards of directors being co-opted by CEOs. Co-opting of powerful ethics offices could similarly facilitate misconduct or other forms of unethical behavior. A co-opted ethics office could facilitate organizational ethics washing by promoting ethical values with respect to some decisions, while at the same time abetting unethical behavior with respect to others. For example, consider an ethics office that ensures that researchers obtain informed consent from research subjects, thus promoting the value of respect for persons, while at the same time using its power to cover up instances of misconduct that, if brought to light, would cost the organizational significantly, both financial and reputationally. In this case, the organization would be engaging in ethics washing via an ethics office that is co-opted, rather than one that is functioning merely symbolically. In short, then, though this article focuses on ethics washing via a merely symbolic ethics office, there are other approaches to ethics washing—some of which are compatible with ethics offices that are not merely symbolic.

3. How Can We Know Whether an Organization's Ethics Office is Merely Symbolic?

This section articulates an assessment framework that can assist in determining whether an AI development organization is engaging in ethics washing. Given our emphasis on ethics offices that are merely symbolic, we focus in this section on what one would need to know to determine whether an organization has a genuine, powerful office of ethics—and hence is not ethics washing via a merely symbolic ethics office. We also, however, provide guidance that can begin to assess whether an organization is ethics washing via other means, such as an ethics office that is under-resourced, selectively committed, or co-opted.[7] The framework consists of (1) structured questions of AI development organizations that researchers, journalists, auditors, regulators, or other investigators might seek to answer, and (2) normative conditionals that can guide the inferences drawn from responses (and non-responses) to these questions.

3.a. Assessment framework: questions

We distinguish between three categories of high-level questions: (1) whether an organization has a genuine ethics office, (2) whether that office has power, and (3) capacities for monitoring and intervention (including capacities possessed by the ethics office, and capacities that others have to monitor the ethics office). We highlight subordinate questions that fall under each of these categories, which can be used to guide inquiry. The first two categories bear on whether the

[7] Schuett, Reuel, and Carlier (2024) provide helpful discussion that is relevant to this section, though we encountered it too late in the publication process to incorporate it fully.

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

organization has an ethics office that is merely symbolic; the third bears on whether the organization might be facilitating ethics washing in some other way. This framework is based on an analysis of the concepts articulated in this article, especially the accounts of ethics offices and organizational power; our interpretation of what information would be required to make informed judgments about whether an organizational subunit is an ethics office and has power, and on standard assessment practices in corporate auditing (e.g., CBN, 2023).

The first category of questions is: Does the organization have a *genuine* ethics office—that is, an ethics office that interacts with other subunits, including R&D subunits, to facilitate the development, implementation, or use of AI in accordance with ethical values? Questions that fall under this category include:

- What are the ethical values or principles that the ethics office promotes? Where are they stated, and to whom? Are they described vaguely and generically, or are they described thoughtfully and in a manner tailored to their specific organizational context?
- What role does the ethics office have in interpreting, operationalizing, and updating the ethical values or principles espoused by the organization?
- How does the organization translate ethical principles or values into practice? Does the ethics office issue specific requirements for other organizational subunits, including R&D subunits, prescribing or proscribing particular design features, data sources, applications, or uses? If so, what are they?
- If the ethics office does issue specific requirements for other organizational subunits, how are these requirements justified? Are they justified by reference to the ethical principles or values that it espouses?
- What specific tasks are performed by members of the ethics office? To what extent are these tasks distinct from the tasks performed by other units?
- How does the organization attempt to ensure that the ethics office can pursue its work without undue constraint or interference by other organizational units and interests?
- Does the ethics office have transparent and confidential channels for reporting ethical problems?

These questions interrogate the content of organizational accounts of ethicality and whether and how they are implemented. If an organizational subunit issues specific and actionable requirements for other organizational subunits with reference to and in consistency with a well-specified set of ethical values, it is, in our usage, a genuine ethics body. Beyond this core account, independence from other organizational subunits and goals, and the existence of structures supporting the reporting of ethical problems, are also indicators of a genuine ethics body. Beyond observation of organizational activities, records of ethics office charter, duties, operations, and interactions with organizational leadership and other subunits may be useful in answering these questions.

The second category is: Does the ethics office have sufficient power within the organization to facilitate the development, implementation, or use of AI in accordance with ethical values? Subordinate questions include:

- How does the organization explain the ethics office's existence?
- What is the structural relationship between the ethics office and the rest of the organization?
- What, if any, adverse consequences would occur for the organization if it did not have a genuine and functioning ethics office?
- What are the formal processes according to which the ethics office makes decisions?
- Are there circumstances under which AI development or implementation projects must receive the ethics body's approval before and regularly during operations? If so, what are these circumstances?

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

- Does the ethics office have the right to refuse approval to projects or require substantial changes for a project to proceed? If so, does it ever exercise this right?
- Do other organizational subunits consistently adhere to the ethics body's requirements or prohibitions?
- How is decision-making in other organizational units affected by the activities and decisions of the ethics office?

There are several signs that an ethics office has power. First, the organization itself may acknowledge that it needs the ethics office to operate; or it may be observable that the lack of an ethics body would result in negative consequences for the organization. An ethics office's power may also be observed in its effects. Instances where it refuses approval to projects, or sets other project requirements, or influences the activities or decisions of other units are also indicators that the ethics office has power. Documentation of ethics-relevant regulatory or other requirements and of implementation of ethics body guidance may be useful for answering these questions.

The third and final set of questions is: What capacities for monitoring and intervention does the ethics office have, and what capacities do others have for monitoring the ethics office? These questions are not necessary to determine whether an organization has a genuine, powerful ethics office. The previous two categories of questions illuminate this. The third category of questions can help to determine whether an ethics office is under-resourced, selectively committed, or co-opted.

- What is the ethics office's budget? Is it sufficient to oversee and intervene upon the organization's AI operations?
- How many members does this ethics office have relative to the organization's AI R&D units? How many are full-time employees? How are they compensated?
- What are the qualifications of the members of the ethics office? Specifically, what qualifications do they have that allow them to identify and apply ethical reasons to organizational activities?
- What information does the ethics body collect about other organizational subunits?
- How is the ethics office's performance assessed? Who assesses it (e.g., other units within the organization, or groups outside of the organization)?
- Who has access to information about the activities and performance of the ethics office, and to whom may this information be communicated? Are current and former employees able to freely criticize the organization on matters of ethics or risk? Does the organization enforce agreements that prohibit or restrict criticism related to ethics or risk?
- How are decisions about initial and continuing membership of the ethics board handled? Can members of the ethics board be fired? If so, under what conditions?

Ideally, an AI ethics body would possess a budget, appropriately skilled human resources, and information-gathering infrastructure sufficient to monitor and intervene upon all an organization's AI activities. Furthermore, there should be mechanisms for monitoring the ethics office and reporting to ensure that it is acting appropriately and that it is not co-opted. Documentation of organizational and ethics office budgets, human resources, and information-gathering procedures may be useful for answering these questions. It is also useful to document whether and how an organization responds to ethics-related concerns by AI researchers and developers, such as the previously discussed open letter by current and former employees of OpenAI and Google DeepMind ("A Right to Warn", 2024).

The assessment framework presented here is not a research instrument or protocol. Rather, it consists (in part) of a structured set of research questions that a researcher, journalist, auditor, regulator, or other investigator could use to frame an investigation of whether a given organization has a merely symbolic ethics office or is engaging in ethics washing by some other means. Such an investigator could search for answers to these questions in a variety of ways. Approaches could include putting them directly to organizational representatives through surveys or interviews, but also including more indirect methods such as documentary review, open or covert observation of organizational operations, or even prodding the organization to see how it responds to ethics-relevant stimuli (e.g., requests for information, expressions of concern, or requests that it fulfill stated commitments).

The questions identified in this subsection are not intended to be exhaustive, and they should be viewed as a starting point for further inquiry; in particular, given the focus of this article on merely symbolic ethics offices, the questions in the third category require further development. These questions in each of these categories can, however, help direct an investigator's attention to information relevant to assessing an ostensible ethics office's genuineness and power. Precisely which signs, and in what combination, are sufficient to conclude that an ethics office is or is not merely symbolic will vary by circumstance, and drawing such a conclusion will require informed judgment. In the next subsection, we will articulate a normative principle that can provide further assistance in guiding inferences from responses (and non-responses) to these questions.

3.b. Assessment framework: normative conditionals

Suppose that an investigator asks an organization about its operationalization of its ethical commitments, for example, in requests for statements, surveys, or interviews. If an organization responds with detailed, thoughtful, and well-articulated responses to the questions, can we reasonably conclude that the organization is not ethics-washing? What if it responds with generic, form-letter answers, such as links to corporate advertising materials? What if it does not respond at all?

If an individual or organization advertises that it is acting in accordance with ethical values or principles, then it has an obligation to respond in good faith to inquiries that critically scrutinize these claims. As noted throughout this article, many AI development firms make public pronouncements that they have ethics offices that facilitate responsible AI; they create press releases that tout their commitment to ethics, justice, and social responsibility, and they develop advertising campaigns intended to convince the public, governments, and others that their products are worthy of trust. When organizations advertise such commitments, they undertake obligations to respond to critical inquiries about their commitments, including inquiries about how their ethics offices actually affect the development, implementation, and use of AI.

If an AI development organization advertises a commitment to ethics based on the operations of an ethics office, and if that organization provides information that evidences favorable answers to the questions in the first two categories, then we have good reason to believe that it has a genuine, powerful ethics office—not a merely symbolic one. Conversely, if it is unwilling or unable to respond to requests for information in a satisfactory manner, then we have good reason to believe that it is engaging in ethics washing—the ethics office that it advertises as being effective is either not a genuine ethics office, or it is one that is merely symbolic or performative.[8]

For our argument, it is important that we are attempting to assess claims about ethics washing, as opposed to ethical behavior more generally. We leave as an open question whether an organization has an obligation to respond to critical scrutiny about its commitments to ethical behavior, policies, and processes if it is not advertising such commitments. If an AI development organization does not signal a commitment to ethical AI, if social scientists or journalists were to ask that organization detailed questions about the structures and processes that facilitate ethical AI, and if that organization refrained from answering those questions, then we leave as an open question what we could justifiably infer on the basis of the organization's non-response. We might have to conclude that we

[8] This conclusion, and the normative conditional that it is based on, is consistent with the view that social science research practices such as surveys and interviews can be seen as interventions, or "field stimulations," and that both responses *and non-responses* to such interventions can count as data that stand in evidential relations to hypotheses (e.g., Salancik, 1979).

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

simply do not know whether it is developing, deploying, or using AI responsibly. But if an organization is signaling a commitment to ethics, then it undertakes obligations to respond to critical inquiries about its commitments. If it fails to respond to such inquiries, we then have reason to believe that it is engaging in ethics washing. This reasoning is defeasible—it could be that the organization is taking ethics seriously, despite the fact that it is failing to respond adequately to critical inquiries. But the failure to respond in good faith to such inquiries constitutes a reason to believe that it is engaging in ethics washing.

4. Scope and Limitations

While our focus in this article is on ethics washing via ethics offices that are merely symbolic, there are multiple ways of engaging in ethics washing. An organization might engage in ethics washing even if it does not have an ethics office. An organization might engage in ethics washing even if it has an ethics office that is not acting merely symbolically. Just as there are other ways besides a merely symbolic ethics office in which an organization might engage in ethics washing, so are there other ways besides the institution of an ethics office by which an organization might attempt to promote responsible design. An organization might, for example, attempt to promote a culture of ethics in which all or most individuals within the organization behave ethically. We focus on ethics offices that are merely symbolic because we believe (again) that this is a common way that AI development organizations engage in ethics washing, even if it is not the only way. Additionally, we hope that our article provides conceptual resources that can be useful in examining ethics washing more broadly, including from organizations with ethics offices that are not acting in a purely performative manner. In Section 3, we identified some other ways in which organizations might engage in ethics washing, and we discussed how the argument of this article is useful in examining them, even though our focus has been on merely symbolic ethics offices.

While the focus of this article is on ethics washing in AI, and while our discussion is motivated by putative examples of ethics washing by AI development organizations, much of our discussion generalizes beyond AI. We focus on AI because many recent accusations of ethics washing have been directed toward AI companies. Additionally, we (the authors of this article) have particular interests in the ethical and responsible design of AI systems and how organizations might facilitate this. However, the argument of this article extends largely, if not completely, to organizations more broadly.

Finally, in addressing the question of how we can assess whether an ethics office is functioning in a merely symbolic or performative manner, the "we" in this sentence refers to individuals who are neither experts in AI nor insiders to the organizations developing AI. It is important for us to be able to make informed assessments of the seriousness with which AI companies take matters of ethics and responsibility. As consumers, we make decisions about which AI systems to purchase or use, as well as how to use them. As citizens, we make decisions about which political representatives and/or positions to support, which in turn have implications for the prospects of AI regulations. In most of these situations, we must make these assessments without the benefit of direct insider knowledge or expertise. This article, we hope, provides some conceptual resources for doing this.

5. Conclusion

Ethics washing in AI is a growing concern. AI development is undertaken primarily by private, for-profit firms, and despite the fact that AI can pose significant risks to the public, there remains a relative paucity of public policies that regulate AI (especially outside of the European Union). Given this, it is especially important that we have resources—both conceptual and empirical—for assessing organizational claims about ethical and responsible AI. This article addresses the question of how we—as outsiders to AI development organizations who are not experts in the technology—can begin to assess whether organizations that advertise commitments to ethical AI are engaging in

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

ethics washing. We provide an account of ethics washing; we identify a common way in which AI development organizations engage in ethics washing—namely via a symbolic or performative ethics office—and we develop tools for identifying symbolic ethics offices in practice. We also articulate other ways of engaging in ethics washing and begin to provide resources for identifying them.

Such investigation is likely to be helpful not only for assessing but also for incentivizing the existence of genuine and powerful ethics offices. Critically investigating claims by organizations that they act in accordance with ethical values is a way of pressuring them to do so. It is also a way of creating the conditions under which ethics offices can have genuine power. For an ethics office to have power, it needs to be responsive to ethics-related internal or external pressure that constitutes a strategic contingency with which an ethics office is especially positioned to cope. Critically investigating organizations about potential ethics washing is one way to create pressure and hold them accountable.

Acknowledgments. Early versions of this article were presented at the University of North Carolina, Chapel Hill; the Southern Workshop in the Ethics of Emerging Technologies (SWEET); the Georgia Tech Philosophy Club, and SPSP 2024. We are grateful to the participants of these events for their feedback. Thanks especially to Joyce Havstad, Carlos Santana, John Walsh, and two anonymous referees for their helpful comments and suggestions. Work on this article is supported in part by the Georgia Artificial Intelligence in Manufacturing Corridor (Georgia AIM), which is funded by the U.S. Department of Commerce Economic Development Administration (EDA) (Grant #04-79-07808). Support for this article has also been provided by the Social Credibility and Trustworthiness of Expert Knowledge and Science-Based Information (SOCRATES), a Kolleg-Forschungsgruppe (Center for Advanced Studies in Humanities and Social Sciences) funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG)—Project 470816212/KFG43. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Georgia AIM, EDA, DFG, or SOCRATES.

Justin B. Biddle is an associate professor in the School of Public Policy at the Georgia Institute of Technology. His research focuses on the role of values in science and technology; the epistemic and ethical implications of the organization of research, and the ethics of emerging technologies, especially artificial intelligence (AI) systems.

John P. Nelson is a postdoctoral fellow at the Georgia Institute of Technology. His research investigates governance of emerging technologies, public impacts of science and technology, and processes of knowledge creation and technology development. He is currently focusing on the implications of artificial intelligence for scientific research and public values.

Olajide Olugbade is a doctoral student of Science & Technology Policy at Georgia Institute of Technology. His research interests include AI policymaking, ethics and governance of emerging technologies, institutional analysis, and responsible innovation.

References

A right to warn about advanced artificial intelligence (2024). https://righttowarn.ai

Ahmad, M. A., Overman, S., Allen, C., Kumar, V., Teredesai, A., & Eckert, C. (2021). Software as a medical device. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, ACM, New York, NY, 4023–4024.

Aitken, M., Ng, M., Horsfall, D., Coopamootoo, K. P. L., van Moorsel, A., & Elliott, K. (2021). In pursuit of socially-minded data-intensive innovation in banking: A focus group study of public expectations of digital innovation in banking. *Technology in Society*, 66, 101666. https://doi.org/10.1016/j.techsoc.2021.101666

Ali, S. J., Christin, A., Smart, A., & Katila, R. (2023). Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. In *2023 ACM conference on fairness, accountability, and transparency*, ACM, New York, NY, 217–226.

Associated Press. (2022). Musk's Twitter has dissolved its Trust and Safety Council. *National Public Radio*. December 12. https://www.npr.org/2022/12/12/1142399312/twitter-trust-and-safety-council-elon-musk.

Biddle, J. (2022). On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, 52(3), 321–341. http://doi.org/10.1017/can.2020.27

Biddle, J. (2023). Organizations and values in science and technology. *Philosophy of Science*. https://doi.org/10.1017/psa.2023.145

Biddle, J., & Kukla, R. (2017). The geography of epistemic risk. In K. Elliott & T. Richards (Eds.), *Exploring inductive risk: Case studies of values in science* (pp. 215–237). Oxford: Oxford University Press.

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. In *ACM FAT* '20: Proceedings of the 2020 conference on fairness, accountability, and transparency*, ACM, New York, NY (pp. 210–219). https://doi.org/10.1145/3351095.3372860

Bürger, V. K., Amann, J., Bui, C.K.T., Fehr, J., & Madai, V. I. (2024). The unmet promise of trustworthy AI in healthcare: Why we fail at clinical translation. *Frontiers in Digital Health*, 6, 1279629. https://doi.org/10.3389/fdgth.2024.1279629.

Central Bank of Nigeria (CBN). (2023). *Corporate governance guidelines*. CBN. https://www.cbn.gov.ng/Out/2023/FPRD/Circular%20and%20Guidelines%20for%20Corporate%20Governance.pdf.

Dahl, R. A. (1957). The concept of power. *Behavioral Science*, 2, 201–215.

De Vynck, G., & Oremus, W. (2023). As AI booms, tech firms are laying off their ethicists. *The Washington Post.* March 30. https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, 67, 559–579.

Edelman, L. B. (1992). Legal ambiguity and symbolic structures. *American Journal of Sociology*, 97, 1531–1576.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16, e12760. https://doi.org/10.1111/phc3.12760

Fisher, E. (2019). Governing with ambivalence: The tentative origins of socio-technical integration. *Research Policy*, 48, 1138–1149. https://doi.org/10.1016/j.respol.2019.01.010.

Hao, K. (2019). *In 2020, let's stop AI ethics-washing and actually do something*. MIT Technology Review. https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/

Havstad, J. (2022). Sensational science, archaic hominin genetics, and amplified inductive risk. *Canadian Journal of Philosophy*, 52(3), 295–320. 10.1017/can.2021.15

Hickson, D. J., Hinings, C. R., Lee, C. A., Schneck, R. E., & Pennings, J. M. (1971). A strategic contingencies' theory of intraorganizational power. *Administrative Science Quarterly*, 16(2), 216–229. https://doi.org/10.2307/2391831

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, G. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, 198, 9941–9961.

Knight, W. (2022). Elon musk has fired Twitter's 'Ethical AI' team. *Wired*. November 4. https://www.wired.com/story/twitter-ethical-ai-team/

Kotliar, D. M., & Carmi, E. (2023). *Keeping Pegasus on the wing: Legitimizing cyber espionage*. Information, Communication & Society, 27, 1–31. https://doi.org/10.1080/1369118x.2023.2245873

Lohne, K. (2021). Ethical capacity and its challenges in the academy of science: Historical continuities and contemporary violence. *Biologia Futura*, 72(2), 155–160. https://doi.org/10.1007/s42977-020-00021-9

McMillan, D., & Brown, B. (2019). Against ethical AI. In *Proceedings of the halfway to the future symposium*, ACM, New York, NY.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83, 340–363.

Microsoft. (2022). Microsoft's responsible AI standard, v2: General requirements. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl?culture=en-us&country=us

Papyshev, G., & Yarime, M. (2022). The limitation of ethics-based approaches to regulating artificial intelligence: regulatory gifting in the context of Russia. *AI & Society*, 39, 1381–1396. https://doi.org/10.1007/s00146-022-01611-y

Piper, K. (2019). Google's brand-new AI ethics board is already falling apart. *Vox*. April 3. https://www.vox.com/future-perfect/2019/4/3/18292526/google-ai-ethics-board-letter-acquisti-kay-coles-james

Roose, K. (2024). OpenAI insiders warn of 'Reckless' race for dominance. *New York Times*. June 4. https://www.nytimes.com/2024/06/04/technology/openai-culture-whistleblowers.html?searchResultPosition=1

Rudner, R. (1953). The scientist *Qua* scientist makes value judgments. *Philosophy of Science*, 20(1), 1–6.

Salancik, G. (1979). Field stimulations for organizational behavior research. *Administrative Science Quarterly*, 24(4), 638–649.

Saltelli, A., Dankel, D. J., Di Fiore, M., Holland, N., & Pigeon, M. (2022). Science, the endless frontier of regulatory capture. *Futures*, 135. https://doi.org/10.1016/j.futures.2021.102860

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, 2(1), 31–42. https://doi.org/10.1109/tts.2021.3052127

Schuett, J., Reuel, A., & Carlier, A. (2024). How to design an AI ethics board. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00409-y

Siapka, A. (2022). Towards a feminist metaethics of AI. In *2022 AAAI/ACM conference on AI, ethics, and society*. ACM, New York, NY.

Schiffer, Z., & Newton, C. (2023). Microsoft just laid off one of its responsible AI teams. *Platformer*. March 13. https://www.platformer.news/microsoft-just-laid-off-one-of-its/

Spiegel, S. (2023). Why salesforce aims to build products that are ethical by design. *Salesforce Newsroom*. January 10. https://www.salesforce.com/news/stories/salesforce-technology-ethics/

Steinhoff, J. (2024). AI ethics as subordinated innovation network. *AI & Society*, 39, 1995–2007. https://doi.org/10.1007/s00146-023-01658-5

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press

Vică, C., Voinea, C., & Uszkai, R. (2021). The emperor is naked. *Információs Társadalom*, 21(2). https://doi.org/10.22503/inftars.XXI.2021.2.6

Westphal, J. D., Gulati, R., & Shortell, S. M. (1997). Customization or conformity? an institutional and network perspective on the content and consequences of TQM adoption. *Administrative Science Quarterly*, 42, 366–394.

Wilson, J., Hume, J., O'Donovan, C., & Smallman, M. (2024). Providing ethics advice in a pandemic, in theory and in practice: A taxonomy of ethics advice. *Bioethics*, 38(3), 213–222. https://doi.org/10.1111/bioe.13208

Wright, J. (2023). The development of AI ethics in Japan: ethics-washing society 5.0? East Asian science. *Technology and Society: An International Journal*, 1–18. https://doi.org/10.1080/18752160.2023.2275987

Cite this article: Biddle, J. B., Nelson, J. P., & Olugbade, O. E. 2025. How Can We Know if You are Serious? Ethics Washing, Symbolic Ethics Offices, and the Responsible Design of AI Systems. *Canadian Journal of Philosophy*: 1–17, doi:10.1017/can.2025.9

https://doi.org/10.1017/can.2025.9 Published online by Cambridge University Press