Digital Society (2025) 4:18
https://doi.org/10.1007/s44206-025-00174-x

ORIGINAL PAPER

AI Ethics' Institutional Turn

Jocelyn Maclure[1] · **Alexis Morin-Martel**[1]

Received: 12 November 2024 / Accepted: 17 February 2025

Abstract

Over the last few years, various public, private, and NGO entities have adopted a staggering number of non-binding ethical codes to guide the development of artificial intelligence. However, this seemingly failed to drive better ethical practices within AI organizations. In light of this observation, this paper aims to reevaluate the roles the ethics of AI can play to have a meaningful impact on the development and implementation of AI systems. In doing so, we challenge the notion that AI ethics should focus primarily on instilling ethical principles in practitioners within AI organizations, as well as the claim that AI ethics can only lead to ethics washing. We propose a two-pronged institutionalist approach to AI ethics, focusing on shaping organizational decision-making processes and emphasizing the necessity of binding legal regulations. First, we argue that AI ethics should give priority to institutional design over the internalization of ethical principles by individual practitioners. We then contend that legally binding rules are needed to this end, both as a motivation for organizations and to contribute to the semantic determination of high-level ethical principles. We then show that promising proposals to operationalize ethical principles require the backing of binding legal norms to be effective. We conclude by highlighting the potential of AI ethics to contribute meaningfully to legislative innovation in AI governance.

Keywords AI ethics · AI governance · Principlism · Ethics washing · Institutionalism · Ethical operationalization

✉ Alexis Morin-Martel
Alexis.morin-martel@mail.mcgill.ca

Jocelyn Maclure
Jocelyn.maclure@mcgill.ca

[1] Philosophy, McGill University, Montreal, Quebec, Canada

Published online: 12 March 2025

1 Introduction

There is growing skepticism regarding the potential of ethical principles to help enact responsible development of artificial intelligence (AI) technologies. In the past five years alone, nearly a hundred different non-legally binding ethical codes or statements have been adopted by public, private, and non-governmental organizations, which put forward mostly the same principles (transparency, fairness, respect for human autonomy, privacy, etc.). However, the concrete effect of these codes on practices has been slow to materialize, and at best modest (Hagendorff, 2020: 108). This alleged inefficacy of 'principlism' led some to argue that the promotion of high-level ethical principles cannot mitigate the serious risks that AI technologies present and is little more than "ethics washing" (Wagner, 2018; Yeung et al. 2020; Bietti, 2020; Van Maanen, 2022; Munn 2023; Steinhoff, 2023; Murgia, 2024). One particularly vexatious instance of ethics washing occurs when the ethical frameworks developed by private companies do not truly aim to guide their practices but rather to enhance their image while weakening the public perception that there is a need for binding laws to regulate their activities (Cole 2022: 2). Ethics washing appears especially worrisome because of the serious ethical concerns that the current fast-paced innovation in machine learning (ML) creates. This is especially true when it comes to the spread of generative tools based on Large Language Models (LLMs) (Kasirzadeh & Gabriel, 2023) and of predictive algorithms used as decision-making tools (Maclure, 2021; Chou et al., 2022; Morin-Martel, 2024) or as recommender systems (Jesse & Jannach, 2021).

While we agree that resorting to non-binding ethical principles is radically insufficient to meet the challenges and risks that the deployment of such systems generates and that ethics washing is a real phenomenon, we believe that critics of AI ethics have an excessively narrow view of what this field has to offer. Critics often reduce the role of ethics in AI to two flawed approaches. The first reduces AI ethics to making professionals and managers within AI organizations[1] internalize ethical principles, virtues, and skills so that they can better assess the relevant ethical challenges that their products raise and (hopefully) mitigate the associated risks. The second one regards AI Ethics as pure ethics washing and window dressing: lofty commitments to values serve to hide ethically tainted practices and goals.

Crucially, we believe that critics of AI ethics neglect the important role that ethics can play when it focuses on institutional design.[2] We suggest that the current state of AI regulation and governance requires adopting an institutionalist view of AI ethics

[1] By AI organizations, we have in mind any entity that is engaged in the research, development, implementation, or application of AI technologies. Such entities can range from private companies to government agencies and public institutions. Additionally, it's important to note that these organizations might not be exclusively or even primarily involved in AI-related activities.

[2] For the sake of this paper, we operate with a loose notion of 'institutions,' which refers to any established framework or system of rules and procedures that facilitates social cooperation and regulates the conduct of individuals within a social sphere or a particular organization. Naturally, it includes public institutions—such as the legal and the political system—which are often the main focus of political philosophers. However, it also includes what Thompson (1999) calls 'midlevel institutions,' i.e. organizations that lie between private life and governmental offices; think, for instance, of "hospitals, schools, corporations, and the mass media" (Thompson, 1999: 110).

that targets AI organizations' decision processes and properly recognizes the role of binding legal regulations as a necessary condition for the responsible development and deployment of AI systems. Our approach is institutionalist in two distinct and mutually reinforcing ways. First, we advocate for the elaboration of ethical frameworks that target the design of decisional processes within AI organizations rather than aiming at making individuals within the industry more ethical or virtuous. Second, we contend that there cannot plausibly be a widespread institutionalization of ethical principles within AI organizations without first having a binding legal framework that mandates it. Therefore, we argue that the most foundational and effective impact AI ethics can currently have is by contributing to the creation of an adequate legal framework for AI.

In Sect. 2, we argue that, to be impactful, ethical frameworks should primarily promote the institutional translation of AI ethical principles at the organizational level rather than internalization at the AI practitioners' level. In Sect. 3, we hold that legally binding legislation is required to specify the content of ethical principles and ensure compliance from AI organizations. To better illustrate this need, we briefly discuss some of the multiple competing metrics that could be used to translate the high-level principle of fairness into an AI credit assessment system. In Sect. 4, we contend that Morley et al.'s (2021) promising "Ethics as a Service" approach is a step in that direction but that they make the crucial error of downplaying the role of hard legal regulations. Finally, this leads us to conclude in Sect. 5 that AI ethics can play a meaningful role in the evolution of existing laws and legislative innovation in uncharted territories.[3]

2 Midlevel Institutions: from AI Professionals' Personal Ethics To Organizational Ethics

In this section, we argue that focusing on individual virtue in AI development is, at best, insufficient and, at worst, integral to an ethics-washing strategy. While it is undoubtedly desirable that individual AI practitioners working within organizations come to internalize relevant ethical principles and virtues, we argue this shouldn't be AI Ethics' priority and main policy. This is because the current structure of AI development is overwhelmingly set within powerful private corporations; a structure that doesn't provide the right conditions and incentives for individual ethical internalization to become a widespread phenomenon. Furthermore, as we will see below, AI practitioners also often lack the proper level of professional autonomy for such an internalization to bear fruit.

Building on Thompson's (1999) institutional turn in organizational ethics, we contend that priority should be given to the fruitful institutionalization of ethical prin-

[3] Although this paper's primary aim is not to flesh out the detailed theoretical view that underpins our practical commitments regarding AI ethics, it's worth pointing out that our thoroughgoing institutionalist approach aligns with a broader metaethical view regarding the most fruitful way to do practical ethics. According to us, a political and legal philosophy-inspired practical ethics is richer and more relevant than the simple application of canonical moral theories (deontology, consequentialism, virtue ethics) to the emerging ethical dilemmas and tensions raised by the omnipresence of AI technologies in human life.

ciples relevant to AI within the organizations that develop and deploy AI systems. Such institutionalization requires AI organizations to set the right kinds of (ethically informed) communicational, deliberative, and decisional processes that favor ethical decision-making. For instance, a corporation working in waste management could be said to have institutionalized certain precautionary principles when their internal policies are designed in a way that minimizes the relevant environmental risks resulting from their activity and makes them proactive when unforeseeable negative consequences emerge as a result of their actions or inaction.

2.1 Internalization of Ethical Principles by AI Practitioners Is not enough

Ethical internalization is often conceived primarily at the individual level. When it is tied to virtue ethics, it refers to the process that leads an agent (plausibly a computer scientist, engineer, or manager) to develop stable dispositions over time to act virtuously (such as becoming courageous, cautious, or benevolent).[4] From that standpoint, developers and managers must acquire the character traits and practical reasoning skills to design truly beneficial technologies and to anticipate (and mitigate) the risks posed by their deployment (Hagendorff, 2022; Neubert & Montañez, 2020). Similarly, there has been significant interest in AI ethics for another form of individual internalization that is more closely associated with deontology called 'principlism.' a popular approach in bioethics that aims at having healthcare professionals internalize certain high-level principles such as non-maleficence, beneficence, respect for autonomy, and justice (Seger, 2022: 45).[5] Although we can hope that as many individual workers as possible in the industry become ethically virtuous developers or that they refer to the right principles in their practice, it would be foolish to rely primarily on this form of internalization in the attempt to regulate AI. To see why this is so, we need to consider some important features of the usual work structure of AI practitioners. For the sake of brevity, we mostly engage with Seger's arguments for principlism in what follows. However, we believe that our arguments also hold against other ethical frameworks that focus primarily on individual ethical internalization, such as Hagendorff's virtue ethics approach.[6]

[4] Alternatively, it could involve coming to see what we identify as moral reasons to be motivating reasons to act (Nagel, 1978).

[5] Although Seger focuses on how principles participate in refashioning the moral outlook of individual practitioners and the ethos of the medical community, Beauchamp and Childress' (2001) version of principlism emphasizes the role of abstract principles (rather than moral theories) in the moral deliberations of healthcare professionals. Rather than simply being applied to particular cases, abstract principles participate in the attempt to reach reflective equilibria between judgments at different levels of generality.

[6] Additionally, we believe that there is a stronger connection than usually assumed between virtue ethics and Seger's version of principlism. Substantiating this claim would go far beyond the scope of this paper, but here is a brief attempt at explaining where the connection lies. Although virtue ethics focuses more explicitly on the development of character traits, it seems highly implausible that one could become a virtuous AI practitioner without referring to certain high-level principles. Similarly, the principlist approach presumably wouldn't produce results that last unless AI practitioners who refer to the four aforementioned master principles ended up internalizing them to a certain extent. Hence, both approaches require individual practitioners to refer to principles and aim to have them develop certain stable dispositions (although these dispositions are probably distinct). More importantly for our current criticism, we see

One *prima facie* argument in favor of pushing for ethical internalization at the AI practitioners' level is that such internalization appears to work in other professional domains. For instance, health professionals seem to have internalized high-level principles such as non-maleficence and respect for patient autonomy. This doesn't mean that all medical practitioners abide by such principles, but they commonly refer to such notions when describing their professional roles and responsibilities as well as to justify their choices. Of course, such high-level principles can be hard to translate into explicit rules and requirements. However, according to Elizabeth Seger, they influence "how practitioners construe the challenges they face and the solutions they entertain" (Seger, 2022: 44–45). For instance, contemporary biomedical ethical principles made the old "doctor knows best" obsolete in favor of emphasizing patient autonomy, and she believes that we could also see a fundamental switch in AI practice (Seger, 2022: 45). A reason to believe that this cultural impact of high-level principles on practitioners is important is that people usually comply more easily with extrinsic rules if they align with their personal values.[7] In essence, Seger argues for a dual approach "in which principlism is supplementary to external policy and recognized as instrumental to the successful implementation of extrinsic rules and regulations" (Seger, 2022: 45). She holds that AI ethicists, such as Mittelstadt (2019), who are skeptical of the impact of such principles, underestimate their cultural influence. Ethical values, after all, contribute to the ethos (or *mores*) of communities. Therefore, she contends that introducing high-level ethical principles in AI development can change how the AI industry behaves.

Seger gives us reasons to believe that the individual internalization of ethical principles by medical practitioners is valuable and has been impactful. However, even if we grant to Seger that it is so, we believe that she neglects how important institutional changes backed by legal sanctions were for these changes of mentality to take place. While medical experts' general shift in mindset regarding medical paternalism is laudable, there are good reasons to believe that it results from the widespread adoption of a series of binding legal rules. For instance, in the past fifty years, new regulations have led to significant changes to the doctors' academic curriculum to emphasize respect for patients' autonomy. Furthermore, the law also began requiring medical care facilities to implement clinical ethics committees dealing with problematic situations, and legal sanctions for doctors who disregard patient autonomy became much more prevalent (Childers et al., 2009; Will, 2011).[8] The evaluative attitudes of physicians did not change spontaneously; court rulings, among other factors,

both approaches as being too narrowly focused on individual practitioners, which makes them susceptible to the objections that we develop in this section.

[7] While Seger does not explicitly mention it, there is empirical evidence regarding that topic in relational theories of procedural justice when it comes to the average citizen's compliance with the law (Gur & Jackson, 2020). Additionally, it seems to us that this need for practitioners to internalize the relevant principles accentuates the connection between principlism and virtue ethics that we briefly sketched in footnote 7.

[8] For instance, the 1972 landmark case *Canterbury v Spence* gave a strong legal foundation to the notion of informed consent in the US, and it played a significant role in changing doctors' behavior. Until then, it was a common practice for doctors not to tell a patient upsetting information, going as far as refusing to let them know that they were dying (Goldie, 1982).

were instrumental. In other words, in the case of medicine, ethical internalization seems to be a consequence of profound changes within medical institutions involving the implementation of rules, procedures, and sanctions for the practitioners who misbehave.

Of course, Seger is not committed to the notion that the promotion of ethical principles *alone* could adequately protect the public against the risks that the deployment of AI systems involves. After all, she is arguing for a dual approach.[9] That being said, the differences between healthcare professionals and AI practitioners' circumstances caution us against relying too heavily on an analogy between them. Indeed, we argue that three fundamental differences should make us doubt that a similar process of ethical learning and internalization is likely to occur in the AI industry in the absence of the institutional turn that we advocate for.

First, AI practitioners typically come from a background in computer science and are not, in most cases, licensed engineers. Employees who are also members of a professional order must often manage a form of dual loyalty. Indeed, they must align with the norms of the organization they work for but also with the deontological constraints that their profession dictates since a failure to do so could put them at risk of losing their right to practice or facing other legal sanctions (Nijhof et al., 2012). Plausibly, this creates a sense of personal accountability that might be fruitful in getting such professionals to internalize relevant ethical principles. However, unlike engineers, who are generally bound by a code of ethics similar to that of medical professionals, programmers who do not hold an engineering license are not subject to the stringent regulations imposed by a professional governing body (Filipovic et al., 2018). If we believe that the adoption of such binding rules was important for the emergence of ethical internalization in the case of medicine, then one has to note that this condition is clearly absent in the case of most AI practitioners. Therefore, this should make us skeptical that a similar internalization can happen.[10]

Second, there are structural reasons to doubt that AI practitioners' widespread internalization of ethical principles could emerge in the first place. Indeed, one chal-

[9] In his own proposal, Hagendorff (2020: 113) also recognizes by the end of the paper the need for institutional changes, such as the adoption of a binding legal framework, in parallel to his virtue ethics approach. However, there seems to be a tension between his main proposal and his later appeal for binding regulations and mandatory institutional measures. Indeed, he notes that his ethical approach aims to move away from a "deontologically inspired tick-box exercise" in favor of pushing AI practitioners to develop certain virtues. He adds that "when following the path of virtue ethics, ethics as a scientific discipline must refrain from wanting to limit, control or steer." He wants "to resign this negative notion of ethics. It should not be the objective of ethics to stifle activity, but to do the exact opposite, i.e. broadening the scope of action, uncovering blind spots, promoting autonomy and freedom, and fostering self-responsibility." (Hagendorff: 2020, 112–113). We are not sure what "ethics as a scientific discipline" entails, but our view is not that an institutionalist and a virtue-based approach are incompatible. Rather, we argue that at this stage in the development and regulation of AI technologies, a normative priority should be granted to the two-tier institutionalist approach that we lay out in this paper.

[10] This isn't to say that we couldn't possibly adopt similar professional ethics frameworks for most AI practitioners. Of course, doing so would necessitate strict regulatory measures, which we firmly support in this paper. Nevertheless, it could be challenging in practice because these workers have diverse professional backgrounds. Additionally, we believe that the other specific challenges associated with AI practitioners mentioned in this section should make us doubtful that merely adopting an AI-related professional ethics framework would suffice for the internalization of relevant ethical principles.

lenge for such an internalization comes from what Thompson (1980) calls the 'problem of many hands.' This refers to the fact that AI development involves multiple actors across multiple domains, which could dilute any sense of personal accountability when negative outcomes occur (Constantinescu et al., 2021: 805). If most AI practitioners do not even understand the full scope of the AI system projects to which they contribute, how could we expect, as Seger would have us believe, that their internalization of certain principles would modify their behavior?[11] This looks pretty different from doctors internalizing certain principles in their day-to-day relationship with their patients. This is not a problem specific to AI, as it occurs in several contexts where responsibility is shared among many actors, such as climate change or the well-known bystander effect (Fischer et al., 2011). Furthermore, unlike doctors, AI practitioners do not have an immediate individual relationship with patients whose interests they must promote (Mittelstadt, 2019). Just as responsibility is diffused among experts, the people who directly suffer harm when AI ethics standards are not upheld are more challenging to identify, further diminishing the perceived moral responsibility of the experts.

Third, even if most AI practitioners internalized the relevant ethical principles, it doesn't mean it would translate into significant changes within the hierarchically organized AI industry. By and large, AI practitioners do not enjoy the same kind of professional autonomy that medical practitioners generally enjoy. Even if they internalized the right principles, for-profit corporations would still have the final say on what gets developed and how it gets designed. Furthermore, as Mittelstadt (2019: 502) argues, going against the company's wishes and acting as a whistleblower often comes at a high personal cost. Even when they do not resort to whistleblowing, it remains risky for employees to push back against their employer's perceived goals. Indeed, very often, there is a significant power asymmetry between employees and employers in private AI organizations. Therefore, even if the employer doesn't directly interfere when employees act contrary to the organization's perceived interest, the mere fact that they *could* do so—for instance, by firing them—limits the actions that employees are likely to take (Pettit, 1997). Because they are subjected to the constant possibility of arbitrary interference, employees lack a particular kind of freedom, which Pettit (1997: 52) calls "republican freedom" and "freedom as nondomination."

This lack of republican freedom also explains why the ethical regulation of AI cannot depend on the creation of "ethical teams" or the appointment of ethics consultants within the industry. For example, it was reported in early March 2023 that Microsoft laid off its entire Ethics and Society Team. This team's goal was to ensure that the high-level ethical principles would be reflected in the design of the products that ended up being commercialized (Steinhoff, 2023). According to credible reports, the group was critical of Microsoft's deployment of new products relying on generative AI, such as the Bing Image Creator, which uses the DALL-E system by OpenAI. This dismissal of the team followed a reminder by Microsoft's VP that the team

[11] Professionals working on large-scale projects can offer meaningful ethical recommendations regarding these projects. However, when they do, it is often in a deliberative setting that includes other stakeholders and employees in a way that allows them to grasp the different stages. This is different from expecting AI practitioners in their day-to-day job, which usually focuses on very narrow tasks, to understand the ethical challenges of their contribution to that project.

should ensure that OpenAI's new models would be in customers' hands as quickly as possible (Schiffer & Newton, 2023). As we can see, even if conscientious developers within a team internalized relevant ethical principles, their ethical concerns within highly competitive capitalist firms such as Microsoft are always likely to take the backseat when they are an obstacle to the firm's economic growth.[12] Furthermore, even if Microsoft had not fired them, the mere fact that they could do so might be enough to deter ethical teams from suggesting significant changes due to concerns about possible retaliation.[13]

In sum, we believe that the lack of hard legal rules and regulatory bodies, the many-hands problem as well as the current general lack of professional autonomy of AI practitioners and AI ethical teams should make us skeptical of the idea that pushing them towards more ethical internalization is the right way to protect the public interest.

2.2 The Institutional Turn: from Virtuous Practitioners To Virtuous Institutional Decision-Processes

A potential solution to this problem, often advocated for in organizational ethics, is to take what Thompson (1999: 109) calls an institutional turn. Institutionalists take seriously the notion that there was a social shift in the relationship between professionals and the organizations they work in. While it was traditionally accepted that they served "their patients and clients in accord with a public-spirited goal," there was a shift in understanding professionals mainly as serving "in organizations that value mainly their expertise and expect them to act in accord with the organization's goal (Thompson, 1999: 109–110). Thompson argues that moral philosophers put much emphasis on ethics at the individual level while, at the other extreme, political philosophers are concerned with much more macro questions such as what should be the basic structure of society and the rights and duties of citizens. In doing so, they both neglect the specific ethical considerations that emerge in what he calls the 'midlevel' institutions, by which he means organizations that lie between our deeply interpersonal private lives and governmental offices such as "hospitals, schools, corporations, and the mass media" (Thompson, 1999: 110). However, the public interacts much more regularly with such institutions than with their government, and such institutions can have a significant impact on their life.

Institutionalists argue that a focus on the personal character and virtue of professionals to safeguard public interest stems from a naïve conception of ethics according to which once ethical principles are correctly established, individuals can simply

[12] Moreover, the fate of the ethical team from Microsoft is not an isolated case. In 2020, Timnit Gebru, who was co-leader of the Google AI ethical team, was forced out of the company because she refused to retract a co-authored paper or to withdraw her name from it (Steinhoff, 2023). The paper pointed out, among other things, the discriminatory risks of large language model systems trained on immense and non-representative training data sets as well as their environmental cost (Bender et al., 2021).

[13] In that sense, even well-designed ethics review boards, as proposed by Schuett et al. (2024), are very unlikely to succeed in that task unless they are embedded in an institutionalist framework such as the one advocated here. Binding regulations protecting the independence of the members of review boards appears crucial to their ability to stand against the organizations when ethical concerns arise.

apply them. This is especially problematic because, insofar as such institutions have duties towards the public, we shouldn't want how a given professional handles hard ethical cases to be settled by their individual morality. It should mostly depend on the organization's policy (Thompson, 1999: 111). In that sense, an approach based on individual morality doesn't sufficiently account for the institutional dynamics and structures within an organization that can either enable or hinder individual ethical behavior. Still, it's important to note that the emphasis on the adoption of better decision processes doesn't mean that we completely move away from individual responsibility. Indeed, the attribution and delineation of individual responsibility in organizations is itself a matter of institutional design. As Thompson notes, managing the many hands problem in complex organizations often requires "…establishing new offices or institutions with individuals specifically charged with overseeing organizational changes to correct structural deficiencies that could result in disastrous failures. Ironically, it requires creating *more* hands—but with more precisely defined responsibilities" (Thompson, 2017: 39).

Going back to AI organizations, we should rightfully fear that AI organizations can weaponize insistence on individual virtue because it suggests that we can rely on such individuals to uphold ethical standards without addressing the collective responsibility and power of organizations to enforce ethical conduct. In short, appealing to (ineffective) individual internalization can allow such organizations to advocate for less stringent rules and for self-regulation. In that sense, a push from the industry for individual internalization might be another sophisticated form of ethics washing. By contrast, AI organizations, through their policies and practices, have the power to affect the development and impact of AI on a much larger scale than individual professionals. They control the resources determining which AI projects are pursued and how they are executed. They also often have complex governance structures that can diffuse responsibility, making it difficult to hold individuals accountable for ethical breaches. Establishing clear ethical frameworks and accountability mechanisms at the institutional level makes it easier to enforce ethical standards in AI development and deployment.[14,15] Therefore,

[14] Like Santoni de Sio and van den Hoven (2018), we place a significant emphasis on the concept of 'meaningful human control' in AI systems development and use, especially when it comes to autonomous or semi-autonomous systems (Maclure, 2021; Morin-Martel, 2024). They argue that to be under meaningful human control, AI systems ought to meet two conditions. First, they must be able to be reasonably responsive to the relevant human developer and users' moral reasons (the tracking condition). Second, they must be designed in a way that allows key human actors to understand "their responsibility for the behavior of the autonomous system" (the tracing condition) (Santoni de Sio and van den Hoven, 2018: 12). While we are sympathetic to their view, our perspective, which is rooted in an institutional view, differs from theirs because we take their position to put too much emphasis on the capacities of individual AI practitioners and on the voluntary adoption of the so-called "Responsible Innovation" program (Santoni de Sio et al., 2024). By contrast, we believe that only an institutional approach that focuses primarily on organizational responsibility can truly provide the foundation for a meaningful human control perspective.

[15] We should note that this paper doesn't aim to provide a full-fledged answer to the "responsibility gap" problem widely discussed in AI Ethics. Broadly, the responsibility gap problem refers to the difficulty of assigning moral responsibility for the actions and outcomes of largely autonomous AI systems because they are not easily traceable or understandable by the humans in charge of designing or putting the systems into operation. However, we think that the institutionalist approach that we defend contributes to the case made by those who believe that AI and new automation technologies do not *necessarily* create a

to be effective, we argue that ethical principles should be instantiated via the AI organizations' decision processes. The institutional turn leads us to ask how the design of an organization's deliberation and decision processes either fosters or hinders ethical behavior. This suggests that to make AI ethics more effective, one must look beyond personal morality and consider how to build responsibility and integrity into organizations themselves. This presumably involves adopting internal rules that force proactive engagement with stakeholders, investment in ethical AI research, and the development of more stringent industry standards. However, this shouldn't be interpreted as a claim that ethical institutionalization is likely to occur within AI organizations if they are not forced to do so. As we will argue in the next section, the ethical internalization we advocate can realistically only be the result of AI organizations complying with binding regulations specific to their sector of activity.

3 The Institutionalization of Ethical Principles by AI Organizations Requires Binding Laws

One challenge we face when trying to push AI organizations to institutionalize ethical principles is to ensure that it does not result in more ethics washing or shallow box-ticking exercises. There are at least two distinct reasons why the meaningful internalization of ethical principles by AI organizations requires the emergence of binding laws specific to AI.

The first is tied to some basic considerations regarding human moral psychology. In the absence of binding legal obligations backed by sanctions, corporate actors lack sufficient incentives to comply with ethical principles. One role of AI laws and regulations is to make it rational for such actors to conform to ethical principles. Of course, this isn't to say that no AI organization would adopt such policies without such constraints. However, given the significant ethical implications and unforeseeable risks associated with the development of AI systems, it would be irresponsible to rely on corporations to act as moral heroes to safeguard public interests. The proliferation of ineffective AI ethics codes within this industry is undoubtedly a good indicator that expecting such moral heroism without binding laws is bound to fail. We side here with business ethicists who think that what matters is compliance with sound democratic regulations rather than empty pledges by corporations to be "socially responsible" (Heath, 2014; Silver, 2021; Bennett, 2023).

While this first problem has been discussed at length, AI organizations also face a second problem in internalizing ethical principles, which is hermeneutical in nature. For AI organizations to internalize ethical principles chosen to regulate their practice, they need the ability to translate them into decision-making processes. In other words, they need the ability to operationalize AI ethics. In the absence of interpretative rules that guide the operationalization of abstract principles, these principles remain severely underdetermined. Thus, even when putting forward principles such

responsibility gap. Indeed, well-designed institutional rules and procedures can allow for the attribution of moral responsibility to specific individual or collective agents for harm done by AI systems. See, for instance, Tigard (2021) on this topic.

as transparency, fairness, and accountability, determining how they should translate into practice often proves difficult.

3.1 The (Limited) Impact of Underdetermined High-Level Ethical Principles on AI Practices

The problem of ethical underdetermination, which can be traced as far back as Aristotle, is about the relationship between abstract moral principles and concrete actions (Aristotle 350 B.C.E./2014: 1140a–b). Ethical principles must guide our actions, but their abstractness requires interpretation to be applied in the often new and unique situations that continuously emerge. In that sense, while general rules are essential, using them aptly in unforeseen situations requires practical wisdom (*phronesis*) to uphold the spirit of the principle, and doing so might even sometimes need exceptions to the rule. One problem specific to AI development is that we don't know how to translate practical wisdom into algorithmic decisions (Wallach & Vallor, 2020: 399). While a system can come to learn to stop a vehicle at a stop sign, it is harder to see how we can imbue it with both common sense and practical wisdom to interpret ambiguous or conflicting signs and signals.

This is especially problematic because AI developers are, in theory, expected to design AI systems that uphold public interests in their decision-making, an orientation often referred to as "responsible AI" or "ethics by design." When AI systems are designed to replace or supplement human decisions and know-how, it appears reasonable to expect such systems' decisions to be compatible with specific values at the core of our democratic societies, such as non-discrimination and respect for human autonomy (Dignum et al., 2018). However, we don't yet know how to develop systems that adequately weigh the ethical considerations tied to their decision-making.[16]

To solve part of the underdetermination problem, AI ethicists have been pushing for a better operationalizing of AI ethics, which requires developing adequate tools to translate high-level principles into practice. Mittelstadt has aptly captured the difficulties that such an operationalization involves. He argues that, unlike medicine, "AI development does not have comparable empirically proven methods to translate principles into practice in real-world development contexts" (Mittelstadt, 2019: 503). He claims that medicine has a set of time-proven institutions and practices (professional orders or societies, clinical ethics boards, deontological codes of conduct, etc.) that help to determine what counts as acceptable day-to-day practice. They help in identifying the hard cases, by flagging what constitutes negligent behavior, and by sanctioning bad actors when the need arises (Mittelstadt, 2019: 503). This set of institutions and processes act as mid-level norms that bridge the gap between high-level principles and low-level requirements of the actual practices (such as clinical judgments). These midlevel norms are themselves grounded on more general healthcare laws and regulations. By contrast, AI lacks such ways to translate principles into practice. For instance, there is no widespread agreement on how to translate a

[16] This is one dimension of the "value-alignment problem" often discussed in AI ethics. Tackling this problem far exceeds the scope of this paper. Russell (2019) as well as Gabriel and Ghazavi (2021) for more on this topic.

principle such as fairness into the decision procedure of an autonomous AI system tasked with rejecting or accepting a credit application (Awwad et al., 2020).[17] This ensures that "conflicting practical requirements will almost certainly emerge across the diverse sectors and contexts in which a principled approach to AI ethics is used" (Mittelstadt, 2019: 504).

Here, we agree with Mittelstadt's verdict. In AI development and deployment, going from principles we want to uphold (such as respect for human autonomy, prevention of harm, fairness, and explainability) to actionable procedures is not straightforward. This has led to the development of many translational tools designed to bridge the gap between ethical principles (the *what*) and actionable procedures (the *how*) (Morley et al., 2021: 239). However, the problem is that multiple competing translational tools can often exist for one principle, and choosing the right tool can be highly context-dependent. To make this complexity more apparent, we turn to a case study by Awwad et al. (2020: 30).

3.2 The Difficulty of Designing the Right Translational Tool

To take but one example, Awwad and his colleagues have shown that multiple concurrent metrics could be used to attempt to ensure that a predictive Machine Learning (ML) system tasked to assess applicants' creditworthiness doesn't discriminate based on race or gender. One way to do so is simply not to provide the system any information regarding the applicants' race and gender. This is known as "fairness through unawareness," which is easy to implement but is problematic because such ML systems tend to use other information (such as graduating from an all-women's college or living in a mostly racialized part of town) as proxies for the characteristics that were meant to be hidden (Awwad et al., 2020: 30).[18]

Because of this well-known problem, many approaches instead tried to tackle fairness issues by providing the ML algorithms with characteristics that are forbidden grounds for discrimination. This is sometimes called "fairness through awareness." Of course, the problem is that there are multiple competing ways to mitigate discrimination through awareness. One is known as "demographic parity," where constraints in the system's optimization are set so that there is a statistical parity among certain pre-defined groups (based on gender and race, for instance) (Awwad et al., 2020: 32). In our credit example, it means that a similar proportion of individuals from all groups who apply should receive loans. However, doing so can significantly impair the algorithm's accuracy because it doesn't consider individual qualifications. Furthermore, just because statistical parity is reached at the group level, it doesn't mean that individuals will not be discriminated against if it is necessary to achieve parity (Cossette-Lefebvre & Maclure, 2022: 1265).[19]

[17] Of course, the lack of agreement about what fairness demands is a broader philosophical problem that isn't specific to AI. However, for the sake of this paper, we focus on a more specific concern for AI ethics that comes from the difficulty of translating ethical principles into practice.

[18] See also Johnson (2021) on the dilemma between reducing bias and maintaining the accuracy of algorithmic predictions.

[19] Additionally, it is worth noting that the very definition of relevant groups is a normative task, notably due to the existence of intersectional experiences (Zimmermann and Lee-Stronach 2022).

In addition, there are two more fine-grained implementations of fairness through awareness known as "equalized opportunities" and "equalized odds." Equalized opportunity refers to the fairness metric that considers only the qualified group. In the context of a credit-scoring system, an equal opportunity model would aim to have the same true positive rate across all groups. This means that among those who are approved for a loan, the percentage of applicants that end up repaying the loan (true positive) needs to be the same, regardless of whether the individual is a member of a protected class or not (Awwad et al., 2020: 35–36). By contrast, equalized odds implement a stronger understanding of the notion of fairness. It requires that both the false negative and true positive rates be equal across all groups. This means that among those who end up being rejected for a loan, there needs to be a similar percentage of applicants who would have ended up paying the loan, had they been accepted (Awwad et al., 2020: 36).[20] The critical difference between the two lies in how they handle negative cases, those who would not repay the loan. Equalized opportunity only ensures fairness among those who get the loan (true positives). However, equalized odds also ensure fairness among those who would have repaid the loan had it been offered to them (false negatives).

To make the distinction between the two approaches more tangible, let us imagine that twenty men and twenty women applied for a loan. Of these twenty men, sixteen men got the loan while twelve women did. Of these sixteen men, twelve ended up repaying the loan, while nine of the twelve women did. Other things being equal, this would pass an equalized opportunities test because the rate of true positive is the same for both groups: 75%. However, we would need to dig deeper to know if it also passes an equalized odds test. Indeed, we would need to see if the disparity between the number of rejected applications between men and women means that more women should have received credit than they did, i.e., we would also need to control for false negatives. While it might seem obvious that equalized odds is a better translation tool for fairness than equalized opportunity, the additional constraints on the algorithm's optimization have been shown to lead to significantly less accuracy. In some cases, it led to a doubling of the loan default rate (Awwad et al., 2020: 37).

The four translational tools presented above only represent a few of the available options in a very narrow kind of operationalization. However, this was meant to show that assessing the right translational tool can be highly complex. The right choice will plausibly depend on what is at stake regarding accuracy and fairness. Sometimes, it might appear perfectly sensible to prioritize the more demanding translation of fairness. Still, there might be other cases where the drop in accuracy could compromise collective goods that we care deeply about. Going back to Seger's argument, we believe this to be an example of the underdetermination of ethical principles in the absence of legally binding norms. Even if AI corporations rightly want to ensure that their AI system does not make decisions based on forbidden grounds of discrimination, there is insufficient legal guidance to reflect how that decision should be opera-

[20] Of course, assessing false negatives is trickier because it requires some probabilistic counterfactual assessment of the sort: if the applicant had been approved for the loan, they would have repaid it. See Awwad et al. (2020) for more on such counterfactual evaluation.

tionalized in the nitty-gritty of contextual decision-making. The principles, therefore, remain severely underdetermined.

This case study illustrates how adopting binding laws and regulations contributes significantly to specifying ethical principles. Indeed, it is no coincidence that fields where ethical internalization of high-level principles works—such as medicine—are strongly regulated by law. As we saw, in the case of medicine, an abstract principle, such as respecting patient autonomy, manifests itself in a series of more explicit rules and procedures to be followed, such as the necessity to obtain free, continuous, and informed consent to proposed care. To guide practice adequately, abstract principles must be complemented by more precise and circumscribed rules that link the general and the particular, including legal norms, professional obligations, and technical standards. Since much remains to be done to update sectorial laws and to pass AI-specific general regulations, it should be unsurprising that we find no widespread agreement in this area regarding the right translational tool to operationalize a high-level principle such as fairness. Moreover, the choice of translational tools and relevant metrics should have a democratic pedigree rather than be left to the whims of private actors.

3.3 The Scope and Specificity Objections

Here, one might worry that we are overestimating the potential of laws and regulations to help solve the underdetermination of high-level AI ethical principles. First, one may suspect that our proposed approach is mistaken insofar as there is no perfect overlap between what is ethical and what is legal (the scope objection). Alternatively, one might wonder whether it is unrealistic to expect the law to be specific enough to provide guidance regarding which translational tool AI organizations ought to use to avoid unfair treatment in each specific context (the specificity objection.) In what follows, we want to briefly respond to these two objections.

Regarding the scope objection, we agree that setting up proper laws and regulations will not in itself address all ethical concerns. Laws establish the foundational baseline for permissible behavior in society. In contrast, ethical values typically encompass a broader set of norms and orientations that guide moral behavior, often exceeding the basic thresholds set by legal norms. This means that even if AI organizations adopted better decision-making policies to comply with new legal obligations, they would nonetheless most probably fall short of other ethical desiderata they should satisfy. For example, critics of generative AI systems have noted that they offer a reductive and one-dimensional representation of various non-Western cultures, such as reducing Islam and Muslim cultures to religious iconography (Qadri et al., 2023: 512). While this constitutes legitimate ethical concern, it seems dubious that there should be a legal obligation to steer clear of stereotypes, unless it can be shown that the expressive acts are hateful and that identifiable individuals are harmed. Still, despite the laws' limited scope, the current gaps in democratic legal regimes regarding AI mean that high-risk AI systems can be deployed with partial impunity (at best). This is why we think that the updating of relevant existing legislation (privacy, consumer protection, liability, intellectual property, free speech, etc.) and the adoption of new AI-specific laws (UE AI Act, Bill C-27 in Canada, etc.) are required and urgent. Although we think that legal evolution and innovation is AI Ethics' current priority,

we agree that this is not a sufficient condition for optimal ethical practices to emerge within the AI industry.

When it comes to the specificity objection, we also agree that expecting laws to cover very fine-grained cases regarding the use of translational tools and procedures is unrealistic. Although they are action-guiding in a much more determinate way than high-level ethical principles, laws of general application also suffer from the underdetermination problem (Aristotle 350 B.C.E./2014: 1140a–b). Nevertheless, we believe that a revised legal framework is a giant step toward solving the underdetermination problem. AI developers and deployers would need to overhaul their operations and design new rules and procedures to comply with the new binding regulations. Moreover, new legal norms can create the right kind of friction in the targeted sector. Indeed, the adoption of new laws and regulations opens the door to litigation regarding their application. Especially in common law systems, judges will play a crucial role in interpreting and applying such regulations to adjudicate individual cases.

To do so, judges often develop and refine legal tests and criteria in their decisions. For instance, in the US, the Americans with Disabilities Act (ADA) provides protection against discrimination in the workspace, forcing employers to reasonably accommodate qualified individuals with disability up to the point of "undue hardship." Of course, what it means for an accommodation to be "reasonable" depends on various factors, such as the operational needs of the business and its financial resources, and several landmark court cases helped interpret that notion contextually over the years (Feldblum, 1996). Similarly, in Canadian law, the legal norm according to which public and private organizations ought to offer reasonable accommodation measures to the members of groups who are disadvantaged by *prima facie* neutral norms is a jurisprudential creation. As we can see, judicial interpretations and the resulting case law can provide more detailed and nuanced guidance than the original legislation itself, and there is no reason to believe that it would be different when it comes to laws that pertain specifically to AI.

It should be clear at this stage that binding legal regulations fulfill two needs: they create the obligation to internalize and institutionalize ethical principles and goals, and they contribute to the semantic determination of abstract principles. However, a third related point worth emphasizing is that they are essential in helping us solve coordination problems. Because they apply uniformly, legal rules establish shared expectations regarding how general duties should be understood in practice and how to settle potential disputes. Knowing how to act often requires settling on a concrete rule that allows everyone to predict the consequences of their actions. For instance, while the legal principle of *fair use* once provided clear guidance on the use of copyrighted content, the rise of generative AI has blurred its boundaries, creating uncertainty about what qualifies as permissible use. In such cases, non-mandatory social norms would likely lack the authority needed to generate a strong basis for stable coordination, even if they could clarify abstract principles.[21]

Of course, it's important to recognize that legal regulations are not a magic bullet because novel cases requiring legal interpretation will always emerge, and there can

[21] We're grateful to the journal's editors for encouraging us to elaborate on this point and for their useful suggestions.

still be a gap between what morality and legality require. Moreover, whenever there is an attempt to create new laws that will affect powerful organizations, we shouldn't underestimate the risk of regulatory capture.[22] Be that as it may, acknowledging this deplorable fact is in no way an argument against legislation. It rather calls for creating the institutional capacity to resist regulatory capture. This is why we believe that laws are required to shape and stabilize the ethical boundaries within which AI organizations ought to operate, even if they are imperfect vehicles to do so and can only set minimal constraints compared to what morality might require.

4 The Ethics as a Service Approach To Operationalization and its Limits

Of course, not everyone agrees that resorting to hard legal governance mechanisms is the best way to solve the underdetermination problem. An example of an ethical operationalization framework that aims to address the underdetermination problem without turning to hard regulations comes from Morley et al. (2021: 242). They believe that hard governance mechanisms suffer from limitations when it comes to the ethical regulation of AI development and deployment. They put forward a specific division of ethical labor that they call "Ethics as a Service," which differs partly from the approach advocated for here with regard to the role and status of hard legal regulations.

In this section, we first give a short presentation of their model and then argue that, while it is a promising one, they fail to draw the right conclusions regarding the essential role of hard regulations for their model to be effective.

4.1 Ethics as a Service

Morley and their colleagues suggest that Ethics as a Service could bridge the gap between the "what" and the "how" in AI ethics, bringing ethical guidance at the design and implementation levels. They argue that operationalizing AI ethics requires the proper division of labor between AI developers and external oversight, ensuring that the ethical constraints for the companies are neither too flexible nor too strict. Indeed, allowing for too much flexibility in the choice of translational tools allows for ethics washing and ethics shopping, which means that companies can pretend to care about ethics by using these tools (ethics washing), or they might select the tools that best fit their interested goals (ethics shopping) rather than what is best for society (Krishnan 2020). By contrast, imposing translational tools that are too strict can be ill-adapted to the specific technologies developed and fail to "account for the fact that sometimes there is no social consensus about what is the 'right' way to interpret or apply ethics or ethical principles" (Morley et al., 2021: 241).

[22] Regulatory capture "is the result or process by which regulation, in law or application, is consistently or repeatedly directed away from the public interest and toward the interests of the regulated industry, by the intent or action of the industry itself" (Carpenter & Moss, 2013: 13).

Morley et al. understand that flexibility can lead to a general lack of accountability from the companies, so they argue for the need for external algorithmic audits (Morley et al., 2021: 248). The role of external ethical auditing is important, but it typically suffers from some limitations. First, they tend to be narrowly focused only on some aspects of the systems, and they happen after the product has already been developed because of the extensive human resources they require. Furthermore, in the context of AI, they often require highly specialized knowledge and are subject to limitations that pertain to trade secrets or the protection of consumers' data (Morley et al., 2021: 248). Finally, Morley et al. argue that another issue with entirely external auditing is that it could end up "ethically desensitizing, de-skilling, and de-responsabilising company employees" and prevent organizations from making "their own critical choices and assume explicit responsibilities" Morley et al. (2021: 249). To help alleviate some of these issues, Raji et al. (2020) suggest that the audits of AI projects be conducted internally within the firms by auditors having full access to the relevant information but who did not work on the project. Of course, internal audits raise another set of issues, such as the lack of incentive for companies to go through a rigorous self-auditing process and the conflict of interests that the auditors might find themselves in (Morley et al., 2021: 249).

Their solution is to implement a "multi-agent system where the responsibility is distributed across different agents (individuals, companies)" and which has the right level of flexibility (Morley et al., 2021, p. 249). Drawing from an analogy with Cloud computing, they argue for what they call "Ethics as a Service,", something similar to "Platform as a service," a middle ground between "Software as a Service" and "Infrastructure as a Service" (Morley et al., 2021: 249).

Software as a Service exemplifies a strictly devolved form of governance of AI. Within this model, all service constituents are managed by a third party, presenting limited opportunities for customization. The analogous AI ethical governance model implies an external entity's dictation of ethical principles, procedures for AI validation, verification, and evaluation stages, and the conduct of an ethical audit, thereby dispossessing AI organizations of any meaningful control.

At the other end of the spectrum, Infrastructure as a Service illustrates an entirely centralized governance model that creates unbridled flexibility. In cloud computing, it involves leaving control over the entire infrastructure to organizations— servers, network operating systems, and storage. Transposing this to AI ethical governance, this model mandates AI organizations to develop AI's ethical principles and procedures internally, with marginal external stakeholder involvement. The onus of conducting internal audits also lies with the AI organizations.

Situated between these two extremes, Platform as a Service embodies a compromise between the devolved and centralized models, operating within an equilibrium of flexibility and rigidity. In this model, while the cloud provider manages core infrastructure components such as operating systems and storage, users are provided a platform to develop custom software or applications. Analogously, in the ethical governance of AI, this implies a balanced model of shared responsibility and control between external and internal entities, paving the way for more comprehensive and collaborative ethical governance.

In practice, Morley and their team suggest that within the Ethics as a Service model, the ethical responsibility is distributed between an external independent multi-disciplinary ethics advisory board and the internal company employees in the following manner. On the external side, the independent multi-disciplinary ethics advisory committee would provide ethical infrastructure (what Floridi (2017) calls the 'infraethics'). It involves three distinct elements. First, developing a principle-based ethical code elaborated through discussions and negotiations while allowing those impacted by the product to have a say in this design. Second, creating a process to be followed at the algorithmic design's validation, verification, and evaluation stages, defining the context-specific meanings of the ethical principles, and offering proven effective translational tools for transitioning from principles to practice. Third, conducting regular audits of the company's overall conduct, not just the end product. This includes assessing the company's commitment to ethical behavior, ensuring that AI organizations adhere to the defined process, and verifying the final output's ethical justification according to contextually defined principles. On the internal side, the company employees would have the responsibility for providing the 'customized software.' This would involve contextually defining the principles developed by the ethical advisory board, identifying the appropriate tools and employing them in designing a specific algorithmic system, and documenting how the process was followed in a public forum (Morley et al., 2021: 251).

4.2 Ethics as a Service Requires Hard Regulations

There is much to be lauded in Morley et al. detailed proposal. Although we are not concerned here with scrutinizing the specific modalities mentioned above, approaches such as Ethics as a Service are steps in the right direction to solve the underdetermination problem. However, we want to raise two issues for their view. The first regards their motivations for turning away from a model of governance that they see as too strict. Our second qualm is that they do not adequately recognize that their demanding governance model can only be impactful if supported by hard regulations that make it mandatory for AI organizations to implement it.[23]

Our first point of contention comes from one of their motivations to avoid an overly top-down (or "strict") approach to ethical operationalization. As stated earlier, they argue that we need to avoid an approach that is too rigid because of a lack of social consensus on how to interpret and apply ethical principles (Morley et al., 2021: 241). While they might be right that there will never be a strong social consensus on the right institutional translation of ethical principles regarding any technological innovation, we believe this is beside the point. In democratic regimes, what matters is democratic lawmaking and inclusive public deliberation (Anderson, 2006: 14), not vague and elusive "social acceptability." A regulatory framework for AI that has a democratic pedigree may be legitimate even in the absence of strong social consensus (Tully, 2008: 147). Similarly, they worry that the imposition of rigid hard regulations

[23] We are not suggesting that all AI organizations should be subject to such a legal obligation to follow the Ethics as a Service model. We agree with Morley et al. (2021: 252) that the risk level that a particular technology represents should be assessed in order to establish the right kind of legal obligations.

on the industry in a top-down manner is unduly paternalistic (Morley et al., 2021: 247). Here again, we think there is nothing paternalistic about submitting an industry to binding regulations that are the result of democratic deliberation and legislation (Shiffrin, 2021). Of course, it is entirely possible that legislators and regulators pass flawed regulations and make wrongheaded decisions, but we assume here that a moderately effective system of checks and balances is in place and that the decisions of public authorities can be ongoingly discussed and criticized.[24]

Turning to our second point of contention, Morley and their colleagues argue that adequately protecting the public interest will require a combination of 'soft' and 'hard' regulations. They endorse the view that "existing 'hard' governance mechanisms (such as legislation and other regulatory frameworks, e.g. ISO requirements) alone provide insufficient protection to individuals, groups, society, and the environment" and that they "do not sufficiently incentivize the design of socially preferable and environmentally sustainable AI" (Morley et al., 2021: 240). According to them, the limitations of current hard regulations explain the turn toward the adoption of 'soft' governance mechanisms, such as "ethics codes, guidelines, frameworks, and policy strategies" (Morley et al., 2021: 240).

To their credit, Morley and their colleagues appear very aware of the failures of grounding soft regulation in highly abstract principles, and they correctly identify the risks of ethics washing and ethics shopping. They also rightly point out that soft regulation mechanisms do not have to remain high level and can offer translational tools that specify abstract ethical principles (Morley et al., 2021: 240). However, we believe they go wrong in their diagnosis of why soft governance mechanisms are ineffective in guiding AI organizations' practices. As we explained earlier, according to them, the culprit is that "almost all translational tools are either too flexible or too strict" (Morley et al., 2021: 241). This is why it is so important to them that we find the right balance in the division of labor between AI organizations and external auditors, a problem that they take Ethics as a Service to help solve.

On our account, the problem is that soft governance mechanisms necessarily need the backing of hard regulations and the legal sanctions they carry. Without such hard regulations, the proper division of labor could be in place, but it is still unlikely that corporations would have the right incentive to comply. As it stands, Morley and their colleagues' position suffers from a mild form of angelism. Indeed, even if major AI organizations adopted their suggested division of labor, this would not guarantee that societal interests would be secured. This is not a criticism of their model per se, which appears to be demanding and sophisticated. Still, we believe that adopting hard regulations that enshrine the legal responsibilities of AI organizations resulting from their Ethics as a Service model, including sanction towards bad faith or negligent actors, is the only way for their proposal to have a meaningful impact.

Indeed, the capitalist market economy in which big AI organizations evolve should make us skeptical that any non-binding ethical processes would be sufficient by themselves to safeguard the interests of the parties affected by AI systems.

[24] Although this is very much a nonideal normative theory paper, we nonetheless rely on a moderately idealized conception of a well-ordered constitutional democracy (Maclure & Weinstock, 2023). https://philpapers.org/rec/MACTCO-77).

This tension between the aspirations of corporations and the operationalization of AI principles led Steinhoff to argue that Morley et al.'s Ethics as a Service doesn't work because they neglect that AI principles will always give way when it conflicts with profit-making. He argues that "the possibilities for the operationalization of AI ethics are predetermined by the requirement that they acknowledge the priority of capital accumulation as a given…so it is to be expected that operationalization does not modify any underlying operations in AI production or deployment" (Steinhoff, 2023). He then concludes that "AI ethics as it stands is a dead-end enterprise. Any interesting AI ethics needs to begin from a perspective which does not prioritize the needs of capital or accept them as given" (Steinhoff, 2023).

Here, we only partially agree with Steinhoff. Hoping that non-binding ethical principles—even when they are broken down into less abstract procedures—will have the desired effect on a sector of industry within capitalist economies appears naïve and severely insufficient, considering what is at stake. However, we do not live in a pure free market capitalistic economy. To different extents, democratic states all intervene in the economy to regulate economic activity and redistribute some of the wealth created. Just as ancient Greek philosophers described political regimes that borrowed from different kinds of constitutions as "mixed regimes," there is a sense in which we live in mixed economic regimes. As such, we should not treat AI any differently than other technologies, such as medical drugs or commercial flights. Indeed, as the political philosophy-inspired approach to business ethics suggests, legislative changes can force the implementation of ethical constraints and processes within AI organizations (Heath et al., 2010). As we alluded to, legal sanctions can make it rational for AI corporations to adopt ethical constraints within a capitalist landscape (Hodges, 2015; Schultz & Seele, 2023). According to our view, the strong incentives for these organizations to comply with the law are what will ultimately lead them to adopt better ethical decision-making processes.[25]

Of course, for this approach to be fruitful, lawmakers need to be convinced that such regulations are required. Furthermore, it is not the case that any regulation will do. Crafting a new legal framework is an especially difficult balancing act. This is why it appears so urgent that AI ethicists who already devoted time and resources to identify the most pressing AI-related ethical issues leverage their expertise to contribute to the creation of the new sectorial and general laws needed to guide AI's development. Here again, it turns out that the preexisting body of work in moral, political, and legal philosophy contains the conceptual resources to understand the distinct and irreplicable contribution of ethical principles in legal innovation.

[25] For instance, as Schultz and Seele (2023) note, the 2001 Enron scandal "led to more formalized implementation processes and legal prescriptions such as the Sarbanes–Oxley Act in the US. In response to the outrage over ethical and financial misconduct, the Sarbanes–Oxley Act legislated ethical behavior for corporations listed on the stock market and their auditors. This included that a code of ethics became a legal requirement for publicly traded companies. Due to the new legislation, business school accreditors began to ask for dedicated business ethics courses and professors that reflect the legal prescriptions" (105).

5 Concluding Remarks: Legal Evolution in Uncharted Territory Requires Ethical Principles

Recently, Luke Munn (2023), a vocal critic of AI ethics, argued that it offers meaningless, isolated principles, and is unable to bridge the gap from principles to practice. Even worse, he holds that the investment of human and financial resources "poured into generating AI ethics frameworks funnels it away from other programs and action" (Munn 2023: 873). It leads him to conclude, rather provocatively, that "it is not enough, then, to denounce AI ethics as fruitless or useless. Instead, a critical assessment of the impact of ethics work to-date must conclude that it is dangerous, hoarding expertise and funding that should be devoted to more effective work" (Munn 2023: 873).[26]

It should be clear by now that we believe Munn's wholesale rejection of AI ethics to be ill-advised and based on an insufficiently rich understanding of the value and role of the ethics of technology. In particular, we think that radical AI ethics critics have lost sight of the role that abstract principles play in the edification of new normative frameworks, and in the evolution of positive law when new social practices reveal its inadequacies. In fact, there are clear signs that we are entering a new phase—a properly legislative one—in the development of governance regimes for AI technologies.

Indeed, it is abundantly clear now that existing legal norms do not suffice to govern and regulate the ubiquitous use of AI technologies. Given the pervasiveness of AI systems in all spheres of contemporary life, laws specifically targeting AI are increasingly seen as essential by governments all over the world. Moreover, laws on matters such as privacy, liability, intellectual property, consumer protection, telecommunication, free speech, and so on, need to be updated in light of the impacts of AI on individual and collective interests. Indeed, there is currently a strong international push toward the adoption of regulations pertaining to AI.[27]

There is great plasticity in the final shape of future AI regulatory frameworks all around the world. This is made evident by the substantial divergence in the legislative approaches adopted by countries where AI systems are predominantly being developed and deployed, such as the US, the European Union, the UK, and China. This divergence in legislative strategies shouldn't come as a surprise. After all, AI is partially uncharted territory, and assessing its potential risks and benefits is highly challenging to legislators because they cannot easily fall back on an existing regulatory framework. In this context, making explicit the ethical problems raised by

[26] This paper offers an indirect response to Munn's challenge. For a more careful analysis of his central claims and a response in defense of AI ethical principles, see Lundgren (2023).

[27] For example, in February 2024, the European Union member countries unanimously endorsed a political deal reached last December regarding the Artificial Intelligence Act (Chee, 2024). In the same vein, Canada is considering passing Bill C-27, a significant portion of which, specifically the third part, is expressly focused on regulating the risks associated with "high-impact AI systems" (Bill C-27 2022). Still, it should be noted that the US appears to be moving in the opposite direction. In October 2023, US President Joe Biden issued an executive order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Biden, 2023). However, in January 2025, the new U.S. President revoked this order through an executive order (Trump, 2025a) and introduced another titled "Removing Barriers to American Leadership in Artificial Intelligence" (Trump, 2025b).

the deployment of new technologies and identifying the values and ends that should guide their design and use are indispensable contributions to their responsible governance. Fortunately, the identification of potential ethical concerns, the elaboration of relevant ethical guidelines, as well as conceptual clarification of central issues in AI (such as the notions of fairness, explainability, etc.) are some of the main contributions from academic research in ethics.

In conclusion, this paper has shown that, contra those who reduce AI ethics to ethics washing, there is a way for ethical principles to help protect society against the risks associated with AI. However, this requires ethicists to acknowledge that AI practitioners' internalization of principles at the individual level won't suffice. Instead, we advocated for a two-tier institutionalist approach that zeroes in on ethical decision-making processes within AI organizations against the backdrop of binding legal regulations. In addition to having a strong motivational force, binding regulations also help to define the practical meaning of these principles. The following discussion of the Ethics as a Service model was meant to show more practically how the operationalization of ethical considerations requires, in our nonideal world, the backing of robust legal regulations. Lastly, the paper also argued that the current stage of AI regulation development presents a unique opportunity for the proposals developed in AI ethics to be impactful on legislation. For this to work, however, AI Ethics must take an institutionalist turn.

Acknowledgements Earlier versions of this paper were presented to the Jarislowsky Chair in Human Nature & Technology and the Research Group on Constitutional Studies at McGill University, at the SSHRC's Insight research group on the legal regulation of AI, and at the ACFAS 2024. We are grateful for the comments from Iwao Irose, Hugo Cossette-Lefevbre, Eran Tal, Natalie Stoljar, Chris Howard, Jacob Levy, Will Roberts, Jordan Walters, Céline Castets-Renard, Jennifer Quaid, and Anne-Sophie Hulin.

Author Contributions Both authors contributed equally to the final paper. The authors' names are to be listed alphabetically.

Funding This research received funding from the Stephen A. Jarislowsky Chair in Human Nature and Technology at McGill University and from an SSHRC Canada Graduate Scholarship.

Data Availability Not applicable.

Declarations

Ethical Approval This work didn't involve any research on human or animal subjects. It is a conceptual philosophy paper.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Competing Interests The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest related to the work submitted for publication.

References

Anderson, E. (2006). The epistemology of democracy. *Episteme*, *3*(1–2), 8–22.

Awwad, Y., Fletcher, R., Frey, D., Gandhi, A., Najafian, M., & Teodorescu, M. (2020). *Exploring Fairness in Machine Learning for International Development* [Technical Report]. CITE MIT D-Lab.

Beauchamp, T. L., & Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bennett, M. (2023). Managerial discretion, market failure and democracy. *Journal of Business Ethics*, *185*(1), 33–47.

Biden, J. (2023). Exec. Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. http://www.govinfo.gov/content/pkg/DCPD-202300949/pdf/DCPD-202300949.pdf

Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 210–219.

Bill, C. (2022). *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, First Session, 44th Parliament.

Carpenter, D., & Moss, D. A. (2013). Introduction. *Preventing regulatory capture: Special interest influence and how to limit it*. Cambridge University Press.

Chee, F. Y. (2024). Europe within reach of landmark AI rules after nod from EU countries. *Reuters*. Retrieved from http://www.reuters.com/technology/france-now-backing-eu-ai-rules-eu-source-says-ahead-bloc-endorsement-2024-02-02/

Childers, R., Lipsett, P. A., & Pawlik, T. M. (2009). Informed consent and the surgeon. *Journal of the American College of Surgeons*, *208*(4), 627–634.

Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, *81*, 59–83.

Cole, M., Cant, C., Ustek Spilda, F., & Graham, M. (2022). Politics by automatic means?? A critique of artificial intelligence ethics at work. *Frontiers in Artificial Intelligence*, *5*, 869114. https://doi.org/10.3389/frai.2022.869114

Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. *Ethics and Information Technology*, *23*(4), 803–814. https://doi.org/10.1007/s10676-021-09616-9

Cossette-Lefebvre, H., & Maclure, J. (2022). AI's fairness problem: Understanding wrongful discrimination in the context of automated decision-making. *AI and Ethics*, 1–15.

Crisp, R. (Ed.). (2014). *Aristotle: Nicomachean ethics*. Cambridge University Press.

Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., & de Wildt, T. (2018). Ethics by design: Necessity or curse? *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 60–66.

Feldblum, C. R. (1996). The (R) evolution of physical disability anti-discrimination law: 1976–1996. *Mental and Physical Disability Law Reporter*, *20*(5), 613–621.

Filipovic, A., Koska, C., & Paganini, C. (2018). *Developing a professional ethics for algorithmists: Learning from the examples of established ethics*. Bertelsmann Stiftung.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*(4), 517.

Floridi, L. (2017). Infraethics–on the conditions of possibility of morality. *Philosophy & Technology*, *30*, 391–394.

Gabriel, I., & Ghazavi, V. (2021). The challenge of value alignment: From fairer algorithms to AI safety. *arXiv preprint arXiv:2101.06060*.

Goldie, L. (1982). The ethics of telling the patient. *Journal of Medical Ethics*, *8*(3), 128–133.

Gur, N., & Jackson, J. (2020). Procedure-content interaction in attitudes to law and in the value of the rule of law. *Procedural Justice and Relational Theory: Empirical, Philosophical, and Legal Perspectives*, *111*.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120.

Hagendorff, T. (2022). A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*, *35*(3), 55.

Heath, J. (2014). *Morality, competition, and the firm: The market failures approach to business ethics*. Oxford University Press.

Heath, J., Moriarty, J., & Norman, W. (2010). Business ethics and (or as) political philosophy. *Business Ethics Quarterly*, *20*(3), 427–452.

Hodges, C. (2015). *Law and corporate behaviour: Integrating theories of regulation, enforcement, compliance and ethics*. Bloomsbury Publishing.

Jesse, M., & Jannach, D. (2021). Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, *3*, 100052.

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961.

Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning Language models with human values. *Philosophy & Technology*, *36*(2), 1–24.

Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, *33*(3), 487–502.

Lundgren, B. (2023). In defense of ethical guidelines. *AI and Ethics*, 1–8.

Maclure, J. (2021). AI, explainability and public reason: The argument from the limitations of the human Mind. *Minds and Machines*, *31*(3), 421–438.

Maclure, J., & Weinstock, D. M. (2023). Two conceptions of public philosophy: A conditional defence of contemporary normative theory. *Civic Freedom in an Age of Diversity: the Public Philosophy of James Tully*, *10*, 25.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Morin-Martel, A. (2024). Machine learning in bail decisions and judges' trustworthiness. *Ai & Society*, *39*(4), 2033–2044.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as A service: A pragmatic operationalisation of AI ethics. *Minds and Machines*, *31*(2), 239–256.

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, *3*(3), 869–877.

Murgia, M. (2024). Overcoming AI ethics, towards AI realism. *AI and Ethics*, 1–6.

Nagel, T. (1978). *The possibility of altruism*. Princeton University Press.

Neubert, M. J., & Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. *Business Horizons*, *63*(2), 195–204.

Nijhof, A., Wilderom, C., & Oost, M. (2012). Professional and institutional morality: Building ethics programmes on the dual loyalty of academic professionals. *Ethics and Education*, *7*(1), 91–109.

Pettit, P. (1997). *Republicanism: A theory of freedom and government*. Oxford University Press.

Qadri, R., Shelby, R., Bennett, C. L., & Denton, E. (2023). AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 506–517.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Santoni de Sio, F., Almeida, T., & Van Den Hoven, J. (2024). The future of work: Freedom, justice and capital in the age of artificial intelligence. *Critical Review of International Social and Political Philosophy*, *27*(5), 659–683.
Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.
Schiffer, Z., & Newton, C. (2023). Microsoft just laid off one of its responsible AI teams. *Platformer*. 14 March. https://www.platformer.news/p/microsoft-just-laid-off-one-of-its
Schuett, J., Reuel, A. K., & Carlier, A. (2024). How to design an AI ethics board. *AI and Ethics*, 1–19.
Schultz, M. D., & Seele, P. (2023). Towards AI ethics' institutionalization: knowledge bridges from business ethics to advance organizational AI ethics. *AI and Ethics*, *3*(1), 99–111. https://doi.org/10.1007/s43681-022-00150-y
Seger, E. (2022). Defence of principlism in AI ethics and governance. *Philosophy & Technology*, *35*(2), 45. https://doi.org/10.1007/s13347-022-00538-y
Shiffrin, S. V. (2021). *Democratic law*. Oxford University Press.
Silver, D. (2021). Democratic governance and the ethics of market compliance. *Journal of Business Ethics*, *173*, 525–537.
Steinhoff, J. (2023). AI ethics as subordinated innovation network. *AI & Society*, 1–13.
Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *American Political Science Review*, *74*(4), 905–916.
Thompson, D. F. (1999). The institutional turn in professional ethics. *Ethics & Behavior*, *9*(2), 109–128.
Thompson, D. F. (2017). Designing responsibility: The problem of many hands in complex organizations. *Designing in Ethics*, 32–56.
Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, *34*(3), 589–607.
Trump, D. J. (2025a). Executive Order: Removing barriers to American leadership in artificial intelligence. http://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence
Trump, D. J. (2025b). Executive Order: Initial rescissions of harmful executive orders and actions. http://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-actions/
Tully, J. (2008). *Public philosophy in a new key: Volume 1, democracy and civic freedom*. Cambridge University Press.
Van Maanen, G. (2022). AI ethics, ethics washing, and the need to politicize data ethics. *Digital Society*, *1*(2).
Wagner, B. (2018). Ethics as an escape from regulation: From Ethics-Washing to Ethics-Shopping. *Being Profiling Cogitas Ergo Sum*, 86–90.
Wallach, W., & Vallor, S. (2020). Moral machines. *Ethics of artificial intelligence* (pp. 383–412). Oxford University Press.
Will, J. F. (2011). A brief historical and theoretical perspective on patient autonomy and medical decision making: Part II: The autonomy model. *Chest*, *139*(6), 1491–1497.
Yeung, K., Howes, A., & Pogrebna, G. (2020). AI Governance by Human Rights–Centered Design, Deliberation, and Oversight. *The Oxford handbook of ethics of AI*, 77–106.
Zimmermann, A., & Lee-Stronach, C. (2022). Proceed with caution. *Canadian Journal of Philosophy*, *52*(1), 6–25.