Discourse Before Doctrine: The Category Error in AI Ethics Education

Lucas J. Wiese and Daniel S. Schiff

Purdue University, USA

This is a draft chapter. The final version is available in Handbook of Critical Studies of Artificial Intelligence and Education edited by Wayne Holmes, published in 2026, Edward Elgar Publishing Ltd, https://doi.org/10.4337/9781035335879

Please cite as: **Wiese, L. J., & Schiff, D. S. (2026). Discourse before doctrine: The category error in AI ethics education. In W. Holmes (Ed.), Handbook of Critical Studies of Artificial Intelligence and Education (pp. 233–246). Edward Elgar Publishing Ltd., https://doi.org/10.4337/9781035335879.00028**

Discourse Before Doctrine: The Category Error in AI Ethics Education

Lucas J. Wiese and Daniel S. Schiff (Purdue University, USA)

Abstract

Centring students in Artificial Intelligence (AI) ethics education is necessary to achieve commonly stated societal desires for responsible, safe, and trustworthy AI. The problem is that AI ethics material is not enough for an educational program to train ethical behaviour in future professionals working with AI systems. Education needs to forego the traditional models of depositing knowledge and skills into students and instead account for the real cognitive-behavioural dispositions of students' lives. We argue that discourse offers a practical method and observable measure for instruction that effectively impacts precursors for behavioural change. Drawing on evidence in cognitive-behavioural psychology, discourse improves the efficacy of decision-making, is normatively tractable, promotes student confidence, and directly qualifies students for AI ethics work. By reframing AI ethics education around discursive practice, we encourage scholars, practitioners, and policymakers to consider the students at the heart of the sociological change expected by the AI ethics field.

Keywords

AI Ethics Education, AI Education, AI Ethics, Responsible AI, Discourse Pedagogy, Moral Psychology

Introduction

This chapter reviews and critiques common practices in the instruction of Artificial Intelligence (AI) ethics. We view AI ethics as an investigation – or at least a useful label – for efforts that minimize the negative impacts of AI and promote positive outcomes. Towards this aim, 'AI ethics education' is one prominent (or at least attempted) strategy. In this chapter, we highlight how educational attempts to promote AI ethics might not lead to the substantive goals of AI ethics very much at all.

The core problem, in essence, is a category mistake. The 'stuff' of AI ethics – its normative principles, practices to safeguard society, and visions of a responsible AI – is not the same 'stuff' that can be dropped into educational curricula that trains students to safeguard and build responsible AI. In other words, placing 'AI ethics' topically into 'education' will not produce the future professionals needed to promote the aims and vision of AI ethics. It is a problem of incommensurability. 'AI ethics' and 'AI ethics education' use the same terminology but cannot be assumed to have the same underlying structure. Thus, despite best intentions, we argue here that the dominant approach to incorporating AI ethics into education will not lead to sufficient student learning or societal change as straightforwardly as is often assumed.

In the following sections, we argue that AI ethics education is not about AI ethics. That is, it is not primarily about the topics, themes, concepts, principles, or even ostensible tools of AI ethics practitioners. The vision of AI ethics – a safe society amidst AI – is constituted of people behaving ethically, rather than unethically, given the plurality and complexity of social AI ethics challenges. Instead, we argue that 'AI ethics education' is about targeting the precursors – the cognitive and behavioural tendencies – of 'ethical behaviour'. Drawing on inspiration and empirical evidence from the broader ethics education literature and when it succeeds or fails, we argue that focusing on these precursors could be far more effective in empowering and encouraging students to meaningfully interpret 'ethical issues', make decisions, and perform ethical actions as future professionals.

Our critical approach challenges the dominant instrumentalist view of AI ethics education that treats AI ethics, often in the form of topics or principles, as mere content to be transmitted to students. By exposing how traditional educational models fail to account for the sociological realities of students – their motivations, behaviours, and real-world professional contexts – we critique the dominant educational structures that produce technically competent but ethically disengaged professionals that ultimately fail to produce safe and responsible AI. We challenge a conception of AI ethics that is superficial and poorly suited to produce the behavioural changes that AI ethics really requires.

In this chapter, we (1) review student-centred considerations from psychology, (2) review AI ethics and educational approaches, and (3) propose 'discourse' as a practical educational solution in response to the cognitive and behavioural limitations of mainstream AI ethics education. We aim to leave readers with practical upshots and recommendations for the future of AI ethics education. We do not attempt to provide a panacea for the problems presented but rather offer just one strategy – one implementable proxy for fostering ethical action in society. To centre students in AI ethics pedagogy, we argue for the benefits of facilitating sound and realistic moral discourse as a method to learn ethical behaviour and measure AI ethics outcomes with respect to cognitive limitations of students.

What's up with students? Education and student experience

> Universities are now largely defined through the corporate demand that they provide the skills, knowledge, and credentials to build a workforce that will enable the United States to compete and maintain its role as the major global economic and military power. (Giroux, 2010, p. 715)

Education has had a conflicting identity for a while. On the one hand, it is commonly understood or argued that it should prepare the next generation of the workforce. That is, it should 'qualify' a student to do a job by arming them with knowledge, skills, and dispositions (Biesta, 2009). These aims are often explicit and organise the curricular program for a field (such as education for the AI ethics workforce), justified by funding, legislation, accreditation, and other dominant systems. However, 'Education' has hidden, and often ignored, effects. Students are implicitly 'socialised' to learn the norms, orders, rules, and principles of good conduct in society, and 'subjectified' such that students ought to become 'subjects' that are free to act on the world, including potentially diverging from socialised norms (Biesta, 2009).

In other words, there are the institutional aims of Education to qualify students, set by policymakers, and the actual (sometimes more mundane) effects of Education experienced by students. We consider how to leverage the socialised and subjectified experience of students to achieve the aim of qualifying students for 'AI ethics' work. To achieve this, "[psychologically]-*wise interventions consider how people's subjective*

meanings exist within a complex system." (Laiduc & Covarrubias, 2022, pp. 223–224). Instead of an educational model where knowledge is merely 'deposited' into students, psychologically attentive approaches appreciate the importance of aligning education with students' life experiences, dispositions, and needs. By no means is this an easy task for educators; But we will draw on critical theory to balance normative theory (the 'oughts' of education) with descriptive-empirical evidence (what happens in education). In doing this, we will assess the standard norms of practice in 'AI ethics education' and raise objections to its shortcomings with the ultimate hope to critically reflect and better realise effective transformations to society at large (Celikates & Flynn, 2023).

A key starting place then understands students and their relationship with ethics and technology. To sketch the psychological and behavioural bounds of students in the AI ethics classroom we will (1) conceptualise relevant social norms and norm psychology to understand the 'socialisation' involved in AI ethics education and (2) present research on the challenges of moral decision-making to explain limitations of current AI ethics education.

Socialisation

To illustrate the process of socialisation, consider some common scenarios that guide our behaviour, such as maintaining an appropriate amount of personal space in a grocery store queue, helping a distressed colleague even if it means working late, or phrasing constructive criticism as a 'compliment sandwich'. These are examples of unspoken rules that we conform to in order to garner acceptance in a group or community (e.g., to be a good citizen, colleague, or mentor). We will understand the effect of these social norms – rules of social behaviour – from a cognitive-evolutionary perspective. Under this lens, we have psychological mechanisms that help us detect 'patterns' of right and wrong, and which allow us to better achieve our individual and communal needs (Kelly & Davis, 2018; Westra & Andrews, 2022). From an evolutionary perspective, recognising and adhering to the community's patterns of right and wrong (i.e., social norms) can reap survival advantage and associated benefits for that community.

Naturally, different people adopt different norms. But at large, people's day-to-day activities are largely governed by what is appropriate or inappropriate in their desired community (Heyes, 2024). Norms are powerful in directing or misdirecting us. Moreover, conforming to norms, and being around other norm-aligned individuals, is metabolically efficient and physiologically comfortable. This 'Sense of Should' instils a subjective pressure to conform to the rules and norms of others (Theriault *et al.,* 2021).

This subjective Sense of Should pressure to conform leads students to adhere to their internalised rules, often resulting in an 'affective friction' when faced with the milieu of new, different, and unfamiliar norms (Kelly & Westra, 2024). This affective friction, when an individual is out of sync with their expected environment, might originate in AI ethics education when the normative recommendations of 'AI ethics' might contrast with the internalised norms and values learned from students' personal experiences. For instance, a computer science student trained to value efficiency might experience affective friction when asked to prioritise fairness over performance optimisation. In turn, when a student recognises mismatches in norms, they will instinctively react with an obstinate resistance to change (Kelly & Westra, 2024). For example, a tech student who values innovation – and sees their professors and peers valuing innovation – may resist 'slowing down' to implement appropriate ethical guardrails. Consequently, as socialised norms are generally powerful in directing or misdirecting us, these realities should be recognised by instructors who expect to teach students the professional practices of AI ethics.

Beyond affective friction, another challenge lies in the largely automatic nature of ethical evaluation. Much of our cognition operates via fast, associative heuristics to conserve energy, often bypassing deliberate and intentional thought (Kahneman, 2003). Indeed, this applies strongly to ethical judgments, where learned socialised rules and patterns become embodied in neurocognitive structures which enables individuals to recognize and evaluate ethical situations without conscious deliberation (Reynolds, 2006). While efficient, this intuitive processing occurs largely outside of our control. Therefore, like anyone else, students make ethical decisions quickly, without much forethought. In sum, affective friction and the automaticity of ethical decision-making

pose problems for instructors who want students to carefully reflect on the ethical decisions they make.

Subjectification

As discussed, automatic intuitive processing primarily leads our ethical decision-making. This poses problems if we want to empower students to be free and prudent actors. Haidt's (2001) pivotal paper argues why 'intuition' bears a primary role and 'reason' a secondary role in judging ethical situations. Meaning, before we become consciously aware of an ethical situation, we have already 'made up our mind' on the matter. Evidence for this 'cognitive unconscious of everyday life' has converged across psychology in the past couple of decades (see Bargh (2022) for a review). While it is undeniably difficult to fully account for intuition in the decision-making process, it is unrealistic to ignore these processes and skip to reason and rationality.

In fact, it is argued that enhanced reasoning skills may lead to negative ethical outcomes. In light of the 2008 US financial crisis, management scholars began to research and find evidence that economics education (often calculated decision-making) encourages self-interested behaviour and greed (Wang *et al.,* 2011). Moreover, Zhong (2011) conducted three experiments that demonstrate when participants were primed to deliberately reason, rather than intuitively reason, they were at least twice as likely to engage in unethical behaviour. Reason and deliberation can certainly help form ethical decisions, but evidence suggests it may not be as strong or universally beneficial as it ostensibly seems.

These findings pose a challenge for ethics education. 'To be effective,' Reynolds argues, "*ethics education should focus on the mental structures of prototypes and moral rules, but it should approach each in different ways.*" (Reynolds, 2006, p. 745). This suggests that rather than solely relying on students' capacity to reason their way to ethical actions, educators should focus on leveraging their pre-existing cognitive dispositions. By helping students engage with their intuitive ethical judgments while simultaneously encouraging them to reform their reasoning, they can begin to

recognise the drivers behind why they do what they do. This process, in turn, helps support their subjectification, empowering them to make more informed decisions.

As Reynolds suggests, tuning into the psychological dispositions and norms of students can more effectively sway pro-ethical behaviour. Here, students may understand more richly what 'influences' their behaviour, and what they can do to 'change' their behaviour amidst the sociopolitical landscapes of organisations and government structures. To hit this on the nose, Murphy (2014) proposes to look directly at the role of intuition such that ethics courses should – at a minimum – teach students about intuitive moral cognition and how to downgrade its effects. In other words, if intuition has such a strong effect, then why not teach students exactly this? The problem with this approach is that intuition is intangible, immediate, and not something that can reliably be measured for improvement. In this chapter, we instead look at a more observable and tangible method to overcome the psychological limitations of ethics education – this being 'discourse' as a minimally viable goal for ethics education. Next, to understand the practicality and feasibility of discourse, we offer a description of the AI ethics field to understand its practices and opportunities.

The substantive field of AI ethics

> The essence of technology is by no means anything technological … Technology is a means to an end [and] technology is a human activity … For to posit ends and procure and utilize the means to them is a human activity. (Heidegger, 1954/1993, pp. 311–312)

Applied in context, AI is merely a tool to procure human needs. Where technology can be seen as an extension of the human body and mind (McLuhan, 1994), we specifically develop and use AI technology to export, facilitate, or ease our cognition or supplement our manual labour with automated systems. Essentially, AI is "*the development of machines capable of learning, reasoning and problem-solving*" (ISO, 2024). Therefore, at its core, given that AI is developed and used for human needs, it follows by logical consequence that there are human, societal, and ethical consequences.

AI ethics, then, is an organised attempt to investigate the impact of AI on humans, or society. We will use Kazim and Koshiyama's definition of AI ethics: "[AI ethics] *is the psychological, social, and political impact of AI.*" (2021, p. 2). This helps understand the broad sweeping scope of AI, not just as the large societal impact and implications in public and private governance, but also at an individual level with psychological impacts amongst others. In this critical view, AI ethics endeavours to balance the protection of individuals with other social (or economic) goals amidst the dissemination of AI systems in society. Succinctly, Waelen provides a useful analysis of AI ethics as a critical theory: "*Both critical theory and AI ethics have a practical goal, namely that of empowering individuals and protecting them against systems of power.*" (2022, p. 11).

To achieve this goal, AI ethics principles (sometimes understood as 'themes' or 'concepts') like transparency, justice, or accountability, amongst others, provide a useful starting point to talk about the pressing societal impacts of AI. Under Waelen's critical orientation, the principle of transparency, for instance, includes the ability for someone to know what happens to their data and how AI affects them. Moreover, justice involves understanding how AI systems and the executives who exert power do not place undue pressure or impose preference on people. Researchers in the AI ethics community have gone to lengths to identify, enumerate, and understand these principles as a proxy for what is in-focus in an AI-infused society. For instance, Schiff *et al.,* (2021) conducted a large global review of AI ethics in the public, private, and non-government sectors, finding relative (at least nominal) importance given to social responsibility, trust, bias and fairness, privacy, and safety. In another large global policy review across sectors, Corrêa *et al.,* (2023) found transparency or explainability, reliability or safety, and justice or equity were at the top of discourse about AI ethics. These concepts have been extensively covered in other studies that conduct meta-analyses on AI ethics principles (to start, see Fjeld *et al.,* 2020; Hagendorff, 2020; Jobin *et al.,* 2019). In sum, these principles set by AI ethics experts aim to provide normative protections for individuals, communities, and society.

The AI ethics field also organises around 'practices' to implement these principles (Kazim & Koshiyama, 2021). Thus, there are 'principles' and there are 'practices', and there are industry, scholarly, and governmental approaches for each. For example, the

workplace can develop policy from principles and implement operational procedures to promote ethics across the AI lifecycle (Prem, 2023) and conduct firm-wide discourse to contextualise ethics (Morley *et al.,* 2021). The research community can review and propose frameworks and tools to implement principles (Lu *et al.,* 2024). And governing or regulating bodies can strategise how to translate ethics into tangible objectives via policy documents and guidelines (Kim *et al.,* 2024). Regardless of 'who' does it, the substantive field of AI ethics has, to a major degree, focused on principles that serve to "*condense 'complex ethical considerations' or requirements into formats accessible to a significant portion of society, including both the developers and users of AI technology*." (Stix, 2021, p. 2).

However, rather than merely the 'complex ethical considerations' common in AI ethics discourse, we ask instead: what are the cognitive, behavioural, and psychological considerations pertinent to AI that actually enable an ethical society? Principles, tools, and frameworks aim to translate abstract concepts into practical concepts, but students become the professionals who will enact those practical concepts. The substantive field of AI ethics rests in the ability to transform principles and practices to action, and that ultimately rests on people. AI ethics' complex ethical considerations therefore need to be translated to complex psychological considerations in the classroom to achieve the substantive goals of AI ethics. Thus, regardless of what students 'can do', a more fundamental question is what they 'will do'. In other words, even if they check the box, recognising all conceptions of AI literacy and AI competency,[1] will they engage in responsible or ethical practices on a day-to-day basis?

Recalling that students' decision-making is shaped by factors like socialised norms, affective friction, and largely unconscious cognitive operations, the limitations of purely procedural, or topical, approaches to AI ethics become apparent. While enumerating principles and practices is important, these alone do not ensure professionals will in fact

[1] For AI literacy we recommend seeing (Chiu *et al.,* 2024; Long & Magerko, 2020; Ng *et al.,* 2021)

adhere to them. Such adherence is fundamentally a matter of moral and behavioural psychology. Thus, a worthwhile task is to pinpoint what 'does' motivate precursors of ethical action in a way that leads professionals to uphold the substantive goals of AI ethics.

Educational approaches around AI ethics

As we reviewed the substantive field of AI ethics, here we will begin to piece together how education wraps around AI ethics. To depict this, we can consider the following metaphor of a crane lowering a rigid 'AI ethics' column into the structure of 'education' without adjusting for the critical pieces of foundation and structural support (i.e., the experiences and cognitive-behavioural drivers of students). Backward design principles are often valued in curriculum and learning: first, identify the intended outcome, then select appropriate instructional modalities for that outcome, and correspondingly choose appropriate material (Wiggins & McTighe, 2005). So, as instructors respond to calls to instil AI ethics skills in students, they teach just that – AI ethics. And regardless of the exact content matter, it is found that instructors maintain flexibility when choosing appropriate pedagogy to elicit desired AI ethics skills (Wiese *et al.,* 2024). In one way or another, topics on AI ethics are taught, and pedagogy promotes AI ethics skills, but along the way, the students' psychological dispositions are overlooked.

We also face challenges of 'how' (or 'where') AI ethics should be taught and 'who' should teach it. Should the material be presented in standalone courses or embedded within existing courses? Should it be taught by technical faculty or ethicists? While educators seem to agree a mix is most effective (Barkhuff *et al.,* 2024), one study indicates that AI practitioners very rarely recognise universities as their source of ethics familiarity (Pant *et al.,* 2024). Moreover, other questions about which competencies to teach, to which grade levels, and to which disciplines make operationalizing AI ethics in education quite difficult (Chiu *et al.,* 2024; Touretzky *et al.,* 2022).

Reviews of AI ethics education attempt to enumerate current approaches, broadly. In a 2020 review of tech ethics syllabi, Fiesler *et al.,* found that topics such as law, privacy,

philosophy, and inequality were the most taught, and that the most intended learning outcomes were to spot issues, critically evaluate them, and make arguments on the ethics topics (Fiesler *et al.,* 2020). And while there appears to be difficulty in aligning learning outcomes with pedagogy (Javed *et al.,* 2022), there are a number of creative and innovative ways to teach AI ethics. For instance, something seemingly simple – making a peanut butter jelly sandwich – can become complicated when children learn about competing stakeholder values (Zhang *et al.,* 2023). These approaches that use hands-on activities and group projects can be useful for students to learn about the risk of AI, biased algorithms, or debates on privacy (Wiese *et al.,* 2024). However, what students learn in the classroom may not transfer to the real-world.

For instance, Tran and Fiesler noticed that while group projects are meant to simulate real-world computing practices, students may not appreciate these connections. Via focus groups with students *in* those group projects, they found that classroom projects were not 'real enough' to truly motivate students. Students do not take ethics seriously in these classroom projects for a simple reason: *because* 'it simply is just a classroom project'. Moreover, the authors find, students do not see the relevance of ethics in their careers, defer the responsibility to someone else, or have not been taught the topic well enough (Tran & Fiesler, 2024). These findings are similar to findings from students in engineering, where ethics is seen as just common sense, not relevant, and not a core subject, among other perceptions that make it challenging for students to begin to even be motivated to seek ethical change (Lönngren, 2021).

In a more optimistic perspective, from a qualitative investigation of 40 AI developers, most do recognise the ethical complexities they inherit as part of their careers (Griffin *et al.,* 2024). And, from a survey of 100 AI practitioners, 87% self-report as somewhat, reasonably, or very familiar with the concept of AI ethics (Pant *et al.,* 2024). However, this familiarity with ethics does not translate to the exercise of ethics in practice. As the practitioners express, the industry that employs them values innovation, and in this, ethics is seen to oppose innovation. Ethical challenges in the workplace take time to consider much less address, and given a lack of AI ethicists trained on these problems, these practitioners have to accept the additional effort it takes to 'do the right thing.' Critically, to solve these challenges, practitioners turn to a key source: each other

(Griffin *et al.,* 2024; Pant *et al.,* 2024). "*In the absence of concrete rules or robust guidance,* developers most often seek help from each other, [and] *this is where they would prefer that ethics conversations occur.*" (Griffin *et al.,* 2024, p. 9, our emphasis). This, then, starts to suggest the power of our proposed minimally viable strategy of discourse as both a method and measure of AI ethics education.

In a review of student reactions and attitudes toward ethical interventions, Padiyath (2024) conducted a synthesised review of what may lead a student to accept or resist ethics in technical computing courses. First, ethics interventions that recognise students' preconceived notions of the world, do not disregard lived experiences, and strategically challenge norms, can lead to positive change. Second, students fear they do not have much agency over ethical change. If an ethics intervention does not contest the power dynamics between *them* and their *workplace*, they will resist ethical change. If students "[lack] *control over their own decision-making in the ethical topic of the intervention, then they will be more likely to resist the ethics intervention*." (Padiyath, 2024, p. 13). In other words, instructors can elucidate the hidden effects of education, namely how socialisation and subjectification occur, to more effectively promote the intentions of tech, computing, or engineering ethics curricula.

In sum, we reviewed how regardless of the educational material or pedagogical implementation, students are typically dissatisfied (or at least unmotivated) with technical ethics instruction. We highlighted that students may accept psychologically wise interventions and that AI professionals may have a natural tendency to seek help from others. To build confidence, our minimal proposal of promoting responsible discourse in the classroom can instil a sense of agency in ethical decisions. As such, next, we will take a look at what discourse might be able to do to target the precursors of ethical action to lead future professionals to uphold the aims of AI ethics.

Discourse

Revisiting Biesta's (2009) concepts of educational qualification, socialisation, and subjectification, we offer discourse as a way to responsibly socialise students, foster their subjective power, and thereby qualify them for exactly what the field of AI ethics

demands. Building on methods like Murphy's (2014) on-the-nose approach to downgrade intuitions, and Padiyath's (2024) 'psychologically realist' recommendations, discourse is something more tangible – something real to enact. Discourse can be seen to incorporate the hidden effects of education and in turn, promote the qualification of AI ethics professionals.

Method

To recap what we previously discussed, ethical action is hard to define and hard to achieve. The source of moral judgment, and motivations for action, stem largely from unconscious processes (Bargh, 2022; Haidt, 2001). For decades, business ethics scholars in particular have been trying to reconcile these challenges to understand the drivers of good and bad decision-making. They ask the underlying question: How do you develop moral behaviour when people inevitably do things they otherwise know are wrong? Theories of moral development have conflicting assumptions and often lack normative implications to develop moral expertise (DeTienne *et al.,* 2021). And it is argued that models of moral automaticity fail to assert educational implications. Meaning, a proposed theory of moral psychology should have the explanatory power to inform educational strategies that can develop or train moral components of the psychological system (Lapsley & Hill, 2008). However, this assertion misunderstands the relationship between descriptive moral psychology and normative educational practices. Rather than attempting to directly alter unconscious and intangible moral constructs (a potentially futile effort), we advocate focusing on observable methods that indirectly influence moral automaticity over time. Thus, what we offer is not an end-all-be-all solution, but rather a method for understanding moral development through promoting and examining discourse amongst students in the classroom, specifically in AI ethics activities.

Naturally, in this proposal, it is important not to overthrow the current methods. As Wiese *et al.,* (2024) found, many current approaches to AI ethics education trend positive. So, we suggest keeping students engaged with the existing case studies, news headlines, and group projects, but to hinge these techniques on a fulcrum of discourse. Get students to truly talk and debate these topics with each other. As a precursor, ask

students what they care about (and carefully listen), then adapt material to make it personally relevant to the students. From this point of personal relevance, interactions in the classroom can connect to students' existing social norms and they can challenge their assumptions against others. If they understand their own positions and their peers' positions, then discourse can promote perspective changing and broader sociotechnical considerations of problems.[2]

This simple – minimally viable – method fits into students' natural propensities to socialise with peers and organically reflect on their moral positions. In response to national and global calls for education on AI ethics, this is what we claim is true AI ethics education. This is what it means to centre students, not topics, and value discourse over doctrine, in AI ethics education efforts. By recognising the lack of power the individual has over their decision-making, we can take a discursive turn to recognize and enable the individual as a part of group decision-making. Next, we will understand why this can be an evaluative measure for a good, qualified, society towards the calls for building safe and responsible AI.

Measure

Discourse, in turn, is not only a method that can be realised in the classroom amidst existing practices but can also be considered a measure of ethical development (at least, one part of ethical development). By assuming discourse as a measure, we implicitly assume that discourse is a 'good' to evaluate (i.e., to value). So, to defend 'discourse' as a good, we will first understand distinctions between collaborativism and individualism, then review the effects of group decision-making, and normative political implications of deliberative democracy. Consequently, given discourse as an evaluative measure, we can not only socialise and subjectify students, but also measure the qualification of students with respect to safeguarding the future of AI development and building responsible AI systems.

---

[2] See Wiese and Magana's (2024) intervention for one example to achieve these aims.

Where 'individualism' assumes optimal human reason as an individual process, 'collaborativism' holds that optimal human reason is socially embedded. Philosopher-neuroscientist John Doris sketches a case against the 'worship of reason' as we are quite ignorant of the sources of our behaviour and action, a perspective well-documented in the literature on moral and behavioural psychology. However, Doris contends that, through collaborative dialogue, we begin to see 'what' we value and 'which' of those values motivate action and behaviour. In this social exchange of values, through dialogue, his minimum claim is simple in that 'two heads are better than one'. When reasoning with others, the shared space of information grows, and individuals have more access to salient pieces of knowledge that they, in turn, provide back to the group. The claim, simply, is that reasoning gets better when it is socially embedded (Doris, 2018).

In decision-making research, there is a well-recognised and modelled 'wisdom of the crowd' effect. The predictions, or judgments, of a crowd are often more accurate than that of even a highly skilled expert. This phenomenon (Surowiecki, 2005) has been mathematically modelled to also show the surprising, but practically important, finding that adding diverse perspectives, rather than skilled perspectives, is more important to accurate judgments (Davis-Stober *et al.,* 2014). These two findings from psychology echo our objectives in this chapter – that discourse amongst peers is an evaluative measure for 'good': good decisions come from good participation. However, 'participation' should not be de facto assumed when invoking the 'crowd', as it does not mean that the crowd is working together. While Tindale and Winget (2019) review this distinction, they nevertheless find that group decision-making still outperforms an individual. While there is in fact danger to loud individual biases swaying a group,[3] the 'shared information' of diverse perspectives has been shown time and time again superior to individual perspectives (Brodbeck *et al.,* 2007; Tindale & Winget, 2019).

[3] As shown in (Gino *et al.,* 2009) where unethical behavior can contagiously spread through group behavior through the perpetuating role of social norms of dishonesty.

The normative upshots of these findings are brought to light in the theory of Deliberative Democracy. It's not just that, statistically, the group outperforms the individual, but that findings from cognitive science inform that there is a 'collective intelligence' (Landemore, 2011) and this provides fertile ground for a valid and legitimate democracy under collective decision-making. First, note that the objective of discourse and deliberation is not consensus.[4] Instead, the objective is to foster critical flexibility where students reconcile old norms with new norms – potentially minimising an otherwise abrasive affective friction (Kelly & Westra, 2024) in the classroom. Discourse is "*no mere abstract demand; rather, it is one which is intimately related to concrete difficulties individuals face in managing contemporary social and cultural pressure*." (White, 1989, p. 81). Discourse helps students interpret and rationalise their life experiences but is also a powerful mechanism toward better decisions.

Deliberative Democracy is empirically found to enable individuals to make sound judgments beyond what would be possible individually. The effectiveness of deliberation sites, such as town halls, citizen forums, or crowdsourcing via social media, is studied to bring about positive change (Dryzek *et al.,* 2019), and we offer education as a hotspot to enable deliberative democracy. As such, classrooms with an ethical focus on discourse can train and foster the safe and responsible development of AI (among other pressing societal issues). Like Arce and Gentile's 'Giving Voice to Values' framework, the intention here is to supplement existing (and practical) material with skills for students to communicate their beliefs, formulate solutions, and transpire value to the organisational development of AI. For instance, students can enumerate various stakeholder parties involved, their values, their arguments and counterarguments, and what levers can invite action (Arce & Gentile, 2015). When students are put in communication with others, the ethical solutions derived through discourse are

[4] Neither is it to change minds, or perspectives, as found in (Wiese & Magana, 2024). Discourse helps students learn, reason, and gain empathy, but does not ultimately impact their decision in a dilemma.

evaluatively more 'good' than in isolation, and consequently, invite large-scale change.

In essence, these arguments establish the qualification effect of discourse when used as a measure of ethical development. Overall, increased discourse is considered an ethically positive outcome, and crucially, in the face of cognitive challenges for individual decision-making, discourse is a vital method for sound ethical decision-making. Students do not resist it, rather, they invite discourse, and it brings engagement and collaboration to the classroom. Furthermore, compared to other proxies for ethical action, discourse is relatively easy to assess through observable indicators. For example: Do students engage in discourse about *X* topic? What is the cogency, empathic perspective taking, or soundness of their discourse? Do students unequally balance emotional motives and logic, or succumb to fallacies? These questions are merely examples of how discourse can be observed and assessed as a measure of ethical development.

Conclusions and Future Work

This chapter argued that AI ethics education, as commonly conceived, often misses its mark. We posited that the substantive field of AI ethics cannot be 'dropped' into an educational curriculum with the expectation of producing the students that the AI ethics field desires. We understood this as the field committing a category error, where AI ethics education has a fundamentally different underlying structure than AI ethics. To substantiate this, we critically examined the cognitive and behavioural dispositions of students, highlighting how pervasive social and moral norms, coupled with limited cognitive control over decision-making, are often overlooked. This examination reviewed the substantive AI ethics field, and how educational instantiations of AI ethics do not appropriately account for the psychological realities of students, and, thus, cannot produce the professionals needed for AI ethics work. Consequently, we demonstrated how discourse-as-a-method can contribute to the proper socialisation of students, subjectification of student agency, and as-a-measure better qualifies students for the real and complex ethical considerations in AI work.

Future research should advance this method-and-measure framework in several directions. First, we need empirical studies that map patterns in discourse to ethical outcomes in professional settings. For instance, which verbal or nonverbal aspects of classroom dialogue correlate with ethical reasoning (or decision-making) quality? How do changes in technological or physical classroom environments affect patterns in discourse that change ethical outcomes? Second, researchers should develop robust metrics to track changes in discourse quality over time. Developing validated assessment techniques would allow instructors to map discourse and moral development in different AI ethics cases.

All in all, the complexity of AI ethics requires more than just knowledge transfer on topics. It requires an aligned educational curriculum that understands the individual 'doing' AI ethics in situated spaces. Building a responsible AI workforce is not just a technical challenge, it is a collaborative challenge that needs interdisciplinary perspectives from cognitive scientists, policy scientists, educational researchers, sociologists, and professional practitioners. By valuing discourse amongst students and professionals, we will ultimately contribute to a more reflective, safe, and equitable future with Artificial Intelligence.

References

Arce, D. G., & Gentile, M. C. (2015). Giving Voice to Values as a Leverage Point in Business Ethics Education. *Journal of Business Ethics*, *131*(3), 535–542. https://doi.org/10.1007/s10551-014-2470-7

Bargh, J. A. (2022). The Cognitive Unconscious in Everyday Life. In A. S. Reber & R. Allen (Eds.), *The Cognitive Unconscious* (1st ed., pp. 89–112). Oxford University PressNew York. https://doi.org/10.1093/oso/9780197501573.003.0005

Barkhuff, G., Borenstein, J., Schiff, D., Uchidiuno, J., & Zegura, E. (2024). Considerations for Improving Comprehensive Undergraduate Computing Ethics Education. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, 1560–1561. https://doi.org/10.1145/3626253.3635557

Biesta, G. (2009). Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability (Formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 33–46. https://doi.org/10.1007/s11092-008-9064-9

Brodbeck, F. C., Kerschreiter, R., Mojzisch, A., & Schulz-Hardt, S. (2007). Group Decision Making under Conditions of Distributed Knowledge: The Information Asymmetries Model. *The Academy of Management Review*, *32*(2), 459–479.

Celikates, R., & Flynn, J. (2023). Critical Theory (Frankfurt School). In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2023/entries/critical-theory/

Chiu, T. K. F., Ahmad, Z., Ismailov, M., & Sanusi, I. T. (2024). What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open*, *6*, 100171. https://doi.org/10.1016/j.caeo.2024.100171

Corrêa, N. K., Galvão, C., Santos, J. W., Pino, C. D., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & Oliveira, N. de. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10). https://doi.org/10.1016/j.patter.2023.100857

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101. https://doi.org/10.1037/dec0000004

DeTienne, K. B., Ellertson, C. F., Ingerson, M.-C., & Dudley, W. R. (2021). Moral Development in Business Ethics: An Examination and Critique. *Journal of Business Ethics*, *170*(3), 429–448. https://doi.org/10.1007/s10551-019-04351-0

Doris, J. M. (2018). Précis of Talking to Our Selves: Reflection, Ignorance, and Agency. Behavioral and Brain Sciences, 41, e36. https://doi .org /10 .1017 /S0140525X16002016

Dryzek, J. S., Bächtiger, A., Chambers, S., Cohen, J., Druckman, J. N., Felicetti, A., Fishkin, J. S., Farrell, D. M., Fung, A., Gutmann, A., Landemore, H., Mansbridge, J., Marien, S., Neblo, M. A., Niemeyer, S., Setälä, M., Slothuus, R., Suiter, J., Thompson, D., & Warren, M. E. (2019). The crisis of democracy and the science of deliberation. *Science*, *363*(6432), 1144–1146. https://doi.org/10.1126/science.aaw2694

Fiesler, C., Garrett, N., & Beard, N. (2020). What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–295. https://doi.org/10.1145/3328778.3366825

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3518482

Gino, F., Ayal, S., & Ariely, D. (2009). Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel. *Psychological Science*, *20*(3), 393–398. https://doi.org/10.1111/j.1467-9280.2009.02306.x

Giroux, H. A. (2010). Rethinking Education as the Practice of Freedom: Paulo Freire and the Promise of Critical Pedagogy. *Policy Futures in Education*, *8*(6), 715–721. https://doi.org/10.2304/pfie.2010.8.6.715

Griffin, T. A., Green, B. P., & Welie, J. V. M. (2024). The ethical wisdom of AI developers. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00458-x

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*(4), 814–834.

Heidegger, M. (1993). Basic Writings: From Being and Time (1927) to The Task of Thinking (1964) (D. F. Krell, Ed.). HarperCollins Publishers.

Heyes, C. (2024). Rethinking Norm Psychology. *Perspectives on Psychological Science*, *19*(1), 12–38. https://doi.org/10.1177/17456916221112075

ISO. (2024, January 31). *What is artificial intelligence (AI)?* International Organization for Standardization. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/news/insights/AI/what-is-ai-all-you-need-to-know.evergreen.html

Javed, R. T., Nasir, O., Borit, M., Vanhée, L., Zea, E., Gupta, S., Vinuesa, R., & Qadir, J. (2022). Get out of the BAG! Silos in AI Ethics Education: Unsupervised Topic Modeling Analysis of Global AI Curricula. *Journal of Artificial Intelligence Research*, *73*, 933–965. https://doi.org/10.1613/jair.1.13550

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

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, *93*(5), 1449-1475.

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. *Patterns*, *2*(9), 100314. https://doi.org/10.1016/j.patter.2021.100314

Kelly, D., & Davis, T. (2018). Social Norms and Human Normative Psychology. *Social Philosophy and Policy*, *35*(1), 54–76. https://doi.org/10.1017/S0265052518000122

Kelly, D., & Westra, E. (2024). Moral Progress is Annoying. *Aeon Magazine*. https://aeon.co/essays/why-does-moral-progress-feel-preachy-and-annoying

Kim, D., Zhu, Q., & Eldardiry, H. (2024). Toward a Policy Approach to Normative Artificial Intelligence Governance: Implications for AI Ethics Education. *IEEE Transactions on Technology and Society*, 1–9. IEEE Transactions on Technology and Society. https://doi.org/10.1109/TTS.2024.3430940

Laiduc, G., & Covarrubias, R. (2022). Making meaning of the hidden curriculum: Translating wise interventions to usher university change. *Translational Issues in Psychological Science*, *8*(2), 221–233. https://doi.org/10.1037/tps0000309

Landemore, H. (2011). Democratic Reason: The Mechanisms of Collective Intelligence in Politics. *Collective Wisdom: Principles and Mechanisms*. https://doi.org/10.1017/CBO9780511846427.012

Lapsley, D. K., & Hill, P. L. (2008). On dual processing and heuristic approaches to moral cognition. *Journal of Moral Education*, *37*(3), 313–332. https://doi.org/10.1080/03057240802227486

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Lönngren, J. (2021). Exploring the discursive construction of ethics in an introductory engineering course. *Journal of Engineering Education*, *110*(1), 44–69. https://doi.org/10.1002/jee.20367

Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *ACM Computing Surveys*, *56*(7), 1–35. https://doi.org/10.1145/3626234

McLuhan, M. (1994). Understanding media: The extensions of man. MIT press.

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. https://doi.org/10.1007/s11023-021-09563-w

Murphy, P. (2014). Teaching applied ethics to the righteous mind. *Journal of Moral Education*, *43*(4), 413–428. https://doi.org/10.1080/03057240.2014.963036

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Padiyath, A. (2024). A Realist Review of Undergraduate Student Attitudes towards Ethical Interventions in Technical Computing Courses. *ACM Transactions on Computing Education*, *24*(2), 1–19. https://doi.org/10.1145/3639572

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in the Age of AI: An Analysis of AI Practitioners' Awareness and Challenges. *ACM Transactions on Software Engineering and Methodology*, *33*(3), 1–35. https://doi.org/10.1145/3635715

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*, *3*(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Reynolds, S. J. (2006). A neurocognitive model of the ethical decision-making process: Implications for study and practice. *Journal of Applied Psychology*, *91*(4), 737–748. https://doi.org/10.1037/0021-9010.91.4.737

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. IEEE Transactions on Technology and Society. https://doi.org/10.1109/TTS.2021.3052127

Stix, C. (2021). Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Science and Engineering Ethics*, *27*(1), 15. https://doi.org/10.1007/s11948-020-00277-3

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Theriault, J. E., Young, L., & Barrett, L. F. (2021). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, *36*, 100–136. https://doi.org/10.1016/j.plrev.2020.01.004

Tindale, R. S., & Winget, J. R. (2019). Group Decision-Making. In R. S. Tindale & J. R. Winget, *Oxford Research Encyclopedia of Psychology*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190236557.013.262

Touretzky, D., Gardner-McCune, C., & Seehorn, D. (2022). Machine Learning and the Five Big Ideas in AI. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-022-00314-1

Tran, M., & Fiesler, C. (2024). "It's Not Exactly Meant to Be Realistic": Student Perspectives on the Role of Ethics in Computing Group Projects. *Proceedings of the 2024 ACM Conference on International Computing Education Research - Volume 1*, 517–526. https://doi.org/10.1145/3632620.3671109

Waelen, R. (2022). Why AI Ethics Is a Critical Theory. *Philosophy & Technology*, *35*(1), 9. https://doi.org/10.1007/s13347-022-00507-5

Wang, L., Malhotra, D., & Murnighan, J. K. (2011). Economics Education and Greed. *Academy of Management Learning & Education*, *10*(4), 643–660. https://doi.org/10.5465/amle.2009.0185

Westra, E., & Andrews, K. (2022). A pluralistic framework for the psychology of norms. *Biology & Philosophy*, *37*(5), 40. https://doi.org/10.1007/s10539-022-09871-0

White, S. K. (1989). The recent work of Jürgen Habermas: Reason, justice and modernity. Cambridge University Press.

Wiese, L., & Magana, A. (2024). Applied Ethics via Encouraging Intuitive Reflection and Deliberate Discourse. *2024 ASEE Annual Conference & Exposition Proceedings*, 46589. https://doi.org/10.18260/1-2--46589

Wiese, L., Patil, I., Schiff, D., & Magana, A. (2024). AI Ethics Education: A Systematic Literature Review. *Working Paper*.

Wiggins, G., & McTighe, J. (2005). What is Backward Design? In *Understanding by Design* (2nd ed., pp. 7–19). Assn. for Supervision & Curriculum Development.

Zhang, H., Lee, I., Ali, S., DiPaola, D., Cheng, Y., & Breazeal, C. (2023). Integrating Ethics and Career Futures with Technical Learning to Promote AI Literacy for Middle School Students: An Exploratory Study. *International Journal of Artificial Intelligence in Education*, *33*(2), 290–324. https://doi.org/10.1007/s40593-022-00293-3

Zhong, C.-B. (2011). The Ethical Dangers of Deliberative Decision Making. *Administrative Science Quarterly*, *56*(1), 1–25. https://doi.org/10.2189/asqu.2011.56.1.001