7

ENGINEERING ETHICS EDUCATION AND ARTIFICIAL INTELLIGENCE

Cécile Hardebolle, Mihály Héder, and Vivek Ramachandran

Introduction

This chapter lies at the intersection of engineering, ethics, education, and artificial intelligence (AI). It discusses how to educate engineers about ethical issues specific to AI engineering and AI *in* engineering, and how AI may be used as a tool in the engineering ethics classroom. As with the other chapters of this handbook, we begin by describing our context, or positionality, as authors.

Positionality

Three academics have written this chapter. The first author, Cécile, is an engineer, computer scientist, and learning scientist working as a pedagogical advisor at École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland. Cécile came late to ethics in the context of her work with teachers. As one of the few women during her engineering and computer science journey, she has been particularly inspired by women in AI ethics. Cécile advocates for practice-oriented, in-context approaches rooted in active and experiential learning.

Mihály is a Hungarian philosopher and computer scientist interested in engineering design, epistemology, and ethics, especially in the context of AI and other software. His career as a software engineer gave Mihály social mobility, a much-needed window to Europe and beyond, and the means to study and teach philosophy at the Department for Philosophy and History of Science at Budapest University of Technology and Economics, which has been his main occupation over the past decade.

Vivek is a non-binary roboticist, learning scientist, and lecturer educated in Asia, North America, and Europe. After completing his Ph.D. in robotics, he shifted focus toward engineering education and ethics based on his desire to emphasize the importance of societal responsibility in engineering. His research explores new ways of teaching ethics to engineers using generative AI as a pedagogical tool; his teaching focuses on developing new curricula for infusing sustainability in all aspects of engineering education.

DOI: 10.4324/9781003464259-9

What do we mean by AI?

The term 'artificial intelligence' has long been a subject of terminological debate. Perhaps the most potent force of canonization was the Russell-Norvig (1995) textbook, which offers a two-by-two matrix of definitions that we summarize thus: AI as relating to internal workings versus observable behavior; AI as performance compared to humans versus an ideal measure.

The lack of total convergence in the definitions is not only a result of the Babelian state of the human race. AI, with its boom-and-bust cycles, can be, at times, an appealing brand, capable of attracting investors and, at the same time, the subject of an increased level of scrutiny, both moral (AI-HLEG, 2019) and legal (Madiega, 2021). Although still under development, the definition we uphold in this chapter is provided by the legal efforts behind the European Union AI Act: "a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments" (Council of the European Union, 2023, p. 29).

Under the hood of AI

The technology that led to the most recent developments in AI is machine learning (ML), through which software can 'learn' from data – particularly non-symbolic ML, such as artificial neural networks. Large language models (LLMs), the technology behind ChatGPT, are a recent evolution of these techniques. From an engineering standpoint, we note that non-symbolic ML generally differs from other types of software or even from older versions of AI:

1. The design process for ML software starts and *centers on data* instead of a set of fixed, human-defined rules.
2. In most cases, the obtained ML model is a *black box***,** and it is hard (if not impossible) to explain how a model produces a given output.
3. The *failure modes* of ML algorithms are significantly different from those of other types of software, making it challenging to ensure the safety and security of ML-based systems.

Although not all AI technologies have the characteristics mentioned above, the ones listed here do generate specific ethical issues that engineers should be able to consider.

Engineers and AI

Concerning AI, we may simplistically consider three categories of roles for engineers: *end-users* (e.g., in AI-assisted engineering); *designers/assemblers* (e.g., designing complex AI systems, embedding AI agents into larger systems such as autonomous vehicles or robots); and *developers* (i.e., implementing AI agents). While some of these roles can be considered the domain of computer science rather than engineering, this distinction is fading as AI spreads across disciplines (e.g., mechanical engineers may contribute to developing AI agents for mechanical applications). This tendency is reflected in the introduction of AI-related courses throughout engineering curricula. Orchard and Radke (2023) report that "the use of AI is pervasive across disciplines such that whether the program majors appear to be AI related is not indicative of their students' engagement with the technology" (p. 15838). As engineering students are increasingly introduced to AI, they should simultaneously be introduced to AI ethics.

Ethical issues specific to AI arise in all of the above-mentioned roles, albeit to different extents. This is why this chapter focuses on *engineers as potential ethical actors in the AI value chain* and discusses the ethical knowledge and competencies engineers need to develop concerning AI.

State of the literature

Ethical questions with AI attract an exponential amount of interest. In December 2023, a Scopus search for 'artificial intelligence' AND 'ethics' returned 5,254 documents and showed that the annual number of publications on AI ethics has been multiplied by ten in just 5 years, from 96 papers in 2017 to 1,000 in 2022. In comparison, scholarship that looked at AI ethics education was much more limited and went from 13 annual publications to 139 over the same period. Strikingly, engineering has not been associated much with this field so far: the annual number of publications found using the query 'artificial intelligence' AND 'ethics' AND 'education' AND 'engineering' was only two for 2017 and 30 for 2022.

In this chapter, we review what exists and where development efforts are needed by considering three main questions: *Where are the ethical challenges for engineers involved with AI? What should engineers know about AI ethics? How can AI engineering ethics be taught, including the use of AI as a tool?*

AI-specific challenges for engineers

Researchers have proposed the notion of 'ethical debt' (Petrozzino, 2021) to refer to the cost generated by negative impacts resulting from ethically flawed systems, in particular AI. This cost is not only borne by system developers, designers, and end-users but also by a range of indirect stakeholders (individuals, communities, societies, and the environment), and it is generally irreversible. Multiple AI-related scandals illustrate how odious that cost may be, such as the thousands of children separated from their families in the Dutch fraud detection scandal (Sattlegger et al., 2022). As potential actors in the decision chain that leads to ethical debt, engineers may face different types of challenges depending on their role.

Engineers as AI users

One frequent claim about AI algorithms is that they can be more 'objective' or 'truthful' than humans. Even a major governmental organization like the Food and Drug Administration (FDA), which plays a major role in pharmaceutical product safety in the United States, suggested in a recent document that AI could "eliminate the subjectivity in the analysis of sophisticated counterfeits" (HHS OCAIO, 2023, p. 3). This widespread belief is contradicted by a large body of research that shows that sources of non-neutrality, subjectivity, and untruthfulness are inherent to the AI production process. For instance, Suresh and Guttag (2021) identified no less than seven different sources of bias throughout the ML life cycle. Worryingly, Griffin et al. (2023) have shown that AI developers also tend to conceptualize AI as value-neutral, with the ethical responsibility lying with the user (an issue we further detail in the following section). This is particularly problematic when AI is used in the engineering design process – ethical flaws in the design tools may induce ethical flaws in the designed products without engineers realizing it. Imagine utilizing an AI-based markerless human pose estimation tool to assess the likelihood of user injury based on the mechanical features of an electric scooter. Contingent upon the dataset it has been trained on, such a tool can be biased (LaChance et al., 2023), and its performance may be lower for specific

user groups. Using such a tool in the engineering process could, therefore, result in serious safety risks for the scooter users.

It is essential that engineers assess ethical risks in AI tools they use – and exercise critical thinking about providers – amid the complex political, ideological, and financial dynamics in the AI field.

Engineers as designers/assemblers or developers of AI

Engineers' responsibility is, of course, more direct in AI designer/assembler or developer roles, where the challenges are also more numerous.

Combined ethical and technical knowledge

In their study of AI developers' agency, Griffin et al. (2023) reported that interviewees described a range of routine technical choices without realizing their ethical dimensions. They suggested that AI developers have "ethical agency 'veiled' as technical agency" (Griffin et al., 2023, p. 6). While this implies that some technical choices in AI entail ethical dimensions, the opposite is also true: some ethical choices in AI entail technical dimensions. For instance, a dedicated field of study researches the fairness of AI algorithms, which has resulted in the development of a range of fairness metrics to assess model fairness as well as technical solutions to try to improve it (Pessach & Shmueli, 2022). As we elaborate later, nearly all dimensions pertaining to the ethics of AI involve some combination of ethical and technical knowledge. Without this combination, engineers will find it challenging to assess and mitigate ethical issues.

Dilemmas

The AI domain is also full of ethical dilemmas disguised as technical dilemmas. Decisions made by AI-powered autonomous vehicles in life-or-death situations, popularized by the Moral Machines project at the Massachusetts Institute of Technology (MIT, n.d.), are a well-known example. Other less visible but perhaps more impactful ethical dilemmas arise in the design decisions engineers make when building AI systems – usually called 'trade-offs' in the AI literature. One example is the fairness–accuracy trade-off: currently, methods that improve the fairness of a model usually decrease its overall accuracy (Pessach & Shmueli, 2022). Many other examples can be found in Sanderson et al. (2023).

Some of these dilemmas not only need recognition and resolution but also re-evaluation. Regarding new technologies, we are usually presented with trade-off situations (Héder, 2021), which often appear to be either using the technology and risking harm or not using it and risking missing out on economic progress. The literature on technological determinism warns us that these first takes are almost always wrong and driven by a misguided 'technological imperative.' Engineers, business owners, and beneficiaries of technological advancements often hastily accept risky features as inherent to technology, implying that society must tolerate these risks. These are false trade-offs, which can be ultimately prevented at marginal, sometimes completely trivial cost – or even no cost at all – with better policies (Héder, 2021, p. 127).

Modularization

Engineers increasingly work with modules that they assemble instead of developing models from scratch (Widder & Nafus, 2022). Generic models such as 'foundation models' can be reused and

fine-tuned for specific applications. Modularity introduces what Widder and Nafus call 'dislocated accountability.' Their interview-based research found that "acknowledgement of harms was consistent but nevertheless another person's job to address, almost always at another location in the broader system of production, outside one's immediate team" (Widder & Nafus, 2022, p. 1). While modularization is not specific to AI, it creates additional challenges in the case of AI because of AI's black-box nature.

Topics in AI ethics

We now present a selection of ongoing conversations in AI ethics that can provide inspiration for AI ethics curricula for engineers. In doing so, we highlight existing controversies and debates, and identify knowledge and skills for engineers to develop. We are not aiming for exhaustiveness in the themes we cover (see Hagendorff, 2022 and Kazim & Koshiyama, 2021 for more complete overviews).

Fairness and bias

Avoiding bias in any kind of system, including AI, is a central concern. It is widely recognized that a biased automated sociotechnical system can cause extreme levels of harm: the well-researched case of the algorithm for analyzing Dutch child benefits (i.e., signaling risks for biased reasons), together with inadequate bureaucratic processes, resulted in tens of thousands of wrongfully canceled child benefit cases (Sattlegger et al., 2022).

Training data is one major source of bias in AI. In applications where generating data for AI training requires some form of human involvement, the process is exposed to cognitive biases. Three are especially prevalent – selection bias, conformity bias, and exposure bias – but there are several more (Chen et al., 2023). Bias can also arise from the model itself, even with unbiased data. For instance, a model may over-generalize from some data points and under-generalize based on others as a result of applying various heuristics that do not have much to do with the semantics of the data. Other sources of bias arise from choices in the model development process (Suresh & Guttag, 2021).

A biased system is unfair, and can take several forms. It may exhibit the Matthew effect, discriminate based on protected attributes (e.g., ethnicity, religion), or exhibit error rates that differ significantly among groups. Current methods to address unfairness issues at the algorithmic level include intervening on the training data, the model, or its output (Pessach & Shmueli, 2022). However, identifying and addressing bias is not always a straightforward statistical exercise. Although some methods can shed light on causal relationships in unfairness issues (Dubber et al., 2020), the definition of bias in certain edge cases requires elaborate philosophical or political discussions (e.g., see Coeckelbergh, 2022, chap. 3, p. 86).

Fairness is among the most widely addressed topics in AI ethics syllabi (Garrett et al., 2020), most frequently introduced through a review of existing fairness metrics with mathematical definitions. While contradictory to each other (Pessach & Shmueli, 2022), these metrics still allow students to perform calculations on example datasets and models and are often used to introduce the philosophical notion of fairness. Fairness evaluation and algorithm auditing are essential skills for engineers to develop – alongside bias mitigation design.

Safety and the alignment problem

Safety is quite a central consideration in AI ethics because of the scale at which these systems can be deployed; even small error proportions can have massive consequences. A significant chal-

lenge is to cope with AI's black-box nature and specific failure modes. Latent errors, hard-to-predict modes of failure, and model drift are examples of the numerous difficulties with safety in AI systems such as self-driving cars (Cummings, 2023). Traditional software safety methods such as testing and code audits are very hard – if not impossible – to use, and public news is rife with examples of AI systems with worrying safety issues, including fatalities (Raji et al., 2022). AI safety risks can be considered at different time scales (Sætra & Danaher, 2023), which is the subject of a raging debate between advocates of the long-term risks – in particular, 'existential risks' (also called 'x-risks') that threaten human existence – and those arguing that more attention should be paid to demonstrated short-term risks that are already affecting populations and the environment.

Autonomy in AI systems, which implies a capacity to make (im)moral decisions, raises a specific safety risk called the 'alignment problem': ensuring that the values manifested in an AI's decisions and acts are aligned with human society. The problem of AI alignment is twofold, according to Gabriel (2020): (1) whose values should an AI be aligned with and (2) how to do the alignment. If the question of selecting the values (to align with) in a pluralistic world is evidently and intrinsically perilous, its implementation is far from trivial as it involves operationalizing the selected values (which will again give rise to debate at another level).

Beyond the performance measures currently central in AI curricula, engineers should be given a practical understanding of AI safety, drawing attention to the potential negative impacts on humans and the environment, both at the micro (individuals) and macro (societies) levels. Evaluating these impacts requires risk assessment methods – an approach also used in the EU AI Act. Finally, an introduction to values and their role in the design process and skills with methods such as value-sensitive design or VSD (Friedman & Hendry, 2019) seem particularly relevant (for more on teaching using such approaches, see Chapter 22).

Transparency and explainability

While traditional AI-leveraged methods were essentially self-explanatory (using logical rules, decision trees, and semantic technologies), these turned out to have less success and more modest capabilities than ML, which, in turn, has a tendency to produce black boxes. An active system that we don't understand – one that makes decisions for us instead of us – naturally raises concerns. The problem is epistemic and the idea is that opacity (Héder, 2023a) takes away our control and our sense of intellectual oversight (Héder, 2023b). On the other hand, transparency can be a way to build trust in the system. The notion of transparency or explainability is, therefore, the most common feature of regulation (Hagendorff, 2020).

However, the fact that explainability and transparency build trust should not be accepted without challenge. Some findings indicate that this effect may present itself only occasionally and may even decrease trust (Scharowski, 2020; Schmidt et al., 2020). Naturally, the transparency of a system to an individual greatly depends on the *a priori* knowledge of that person about how AI works, as well as the person's level of exposure to the system. Therefore, the draft standard in this question (P7001, Winfield et al., 2021) distinguishes between expert, user, and bystander roles.

Human agency

The increasing presence of AI systems in our lives raises questions regarding our agency (Prunkl, 2022): *Is our agency augmented or antagonized by increasing AI autonomy?*

AI's impact on human agency is a double-edged sword. While AI tools have contributed to improving people's quality of living, often by employing their data to provide tailored recommendations (Logg et al., 2019), there are concerns about how the data is obtained, stored, and utilized – and controversies regarding manipulation and surveillance (Floridi et al., 2021; Ienca, 2023). AI tools like chatbots may seem impressive at mimicking human interactions that seemingly display feelings of empathy (Stark & Hoey, 2021), and that characteristic can add to the automation bias problem – where humans overly trust AI recommendations – that undermines critical thinking and accountability (Ienca, 2023; Suresh et al., 2020). Moreover, AI can also perpetuate falsehoods and contribute to the illusory truth effect (i.e., the propensity for humans to believe misinformation as truth by dint of repetition).

This highlights the ethical responsibility in engineering to engage (as users and creators) with AI systems in a way that respects user boundaries, maintains transparency, and upholds ethical interaction standards. A balanced approach is essential in classroom discussions, examining the potential benefits *and* the ethical issues AI systems pose – as this will determine how human empowerment and agency are protected and strengthened. For a more detailed critique of how AI autonomy affects human agency, we refer readers to Mhlambi and Tiribelli (2023).

Sustainability

Currently vastly under-addressed in typical AI curricula, sustainability questions materialize a central dilemma: AI offers some potential for addressing some of the complex climate change issues (Larosa et al., 2023) while at the same time requiring colossal amounts of resources, including energy, data, hardware, and human labor (Bender et al., 2021). The complex cost–benefit questions related to AI should not be left out of current efforts to introduce sustainability into engineering programs.

While the environmental impacts of AI in general remain massively undocumented, recent studies on LLMs tend to show that both the carbon and the water footprints of these systems are significantly larger than for other IT systems (Li et al., 2023; Luccioni et al., 2023; Patterson et al., 2021). In addition to parameters related to cloud infrastructure, the size of the datasets and models, but more importantly their architecture, seem to increase the impact at the time of both training and use. The GPT models (e.g. ChatGPT) seem to have a particularly high environmental impact, which is concerning given the attention they generate in (engineering) education.

Unfortunately, AI also presents other sustainability issues (Crawford, 2021). Researchers have investigated the questionable labor practices behind AI (Hagendorff, 2022), exemplified by the Kenyan workers who made ChatGPT less toxic, reducing the amount of violent, racist, and sexist outputs for end-users by reviewing and labelling harmful content manually. On the hardware side, although many sustainability issues are not AI-specific but cloud-computing related, the exponential increase in dataset and model size leads to a race for optimized hardware. In addition to the catastrophic environmental impact of hardware production (Crawford, 2021), these impact reduction efforts are likely to be counteracted by increasing demand (rebound effect, see Grubb, 1990).

Although more research is needed, engineers should be introduced to these issues as early as possible and develop skills for evaluating AI systems' carbon and water footprints. The systemic nature of these issues also calls for macro approaches encompassing the whole AI life cycle and including questions of resources and labor dynamics at a large scale. In particular, engineers need to develop systems thinking skills and practice with methods such as life-cycle assessment.

Regulation of AI

The successes of AI in the 2020s provoked a wave of soft laws and regulations (Héder, 2020). One major challenge is assigning responsibility for unfortunate or unwelcome events. Since responsibility is closely associated with decision-making, AI automated decision-making has created a responsibility gap (Matthias, 2004): a system making a decision is not a legally accountable agent, unlike the human being it replaced. Therefore, responsibility needs to be assigned elsewhere, but this redistribution is far from trivial. Another issue is the vast potential of AI for technology lock-in because, as with any software, once developed at significant expense, the margin cost of reproduction is minimal. The fact that software can be reused cheaply and infinitely removes the incentive for creating a completely new one at high capital expenditure. The lack of serious new computer operating system projects illustrates this point quite well. In this case, the decisions made in the early stage, lacking information and foresight, may have long-lasting consequences. Finally, generative AI challenges existing copyright and intellectual property frameworks.

In addition to theoretical background – on how norms are created or studying certification materials and reports of actual systems – mock evaluation sessions and simulated certification processes (e.g., where one team of students act as the product owners while others as the certifying body) can provide engineers with a pragmatic understanding of regulatory issues. Yet, the rapid evolution of AI regulations will make it challenging to keep educational material up to date.

Pedagogical methods

We now turn to the pedagogical methods that could be used to teach engineers about AI ethics. Although still few, there are some reviews of AI ethics syllabi, mainly in the United States (Garrett et al., 2020; Raji et al., 2021; Saltz et al., 2019; Tuovinen & Rohunen, 2021). They tend to show that the range of pedagogical methods used in AI ethics is quite diverse and has much in common with engineering ethics education (EEE) methods – the overall topic of this handbook. The following subsections provide an overview of existing approaches and identify avenues for future work related to EEE and AI. We will discuss the specific case of how AI could be used as a tool for EEE.

General engineering ethics methods

Readings followed by class discussions are among the most frequently used methods in AI ethics classes (Garrett et al., 2020; Raji et al., 2021; see also Chapter 25 on reflective and dialogical approaches to teaching EEE). Reading lists generally include research papers and news articles that help relate course content to current events. Although academic readings may provide insights into the multidisciplinary nature of AI ethics (Raji et al., 2021), the vocabulary used may create difficulties for students, and engineering students generally have little experience with these methods, especially at the undergraduate level (Tuovinen & Rohunen, 2021).

Pedagogical methods in AI ethics also include *case studies* (e.g., see Alam, 2023; and Chapter 20) for students to practice assessing ethical dilemmas and ethical decision-making. Unfortunately, there's no shortage of opportunities to build cases on AI-related real-world events. Cases are frequently used with other techniques such as *role plays* (Hingle & Johri, 2023; and Chapter 24) and *debates* (Alam, 2023; and Chapter 25). Some AI ethics courses make use of *science fiction* (Burton et al., 2018; and Chapters 13 and 24) to equip "students with skills to cope also with the unforeseen ethical issues in their future work" (Tuovinen & Rohunen, 2021, p. 21). *Games* are another

pedagogical tool used in AI ethics, whether in digital or physical form (Alam, 2023; Hardebolle et al., 2022).

Experiential practice-based approaches

Practice-based approaches are a central component of AI courses for engineers. Exercises and projects (see Chapter 21 on problem-based learning in EEE) provide opportunities to experience the AI development process, which, besides aiding future AI developers, also enhance engineers' understanding of the underlying mechanisms of AI. Below, we discuss how such activities can provide opportunities to teach and learn AI ethics in context.

Exercises with data

The data on which AI systems rely provides valuable opportunities for ethics education. Public datasets such as the COMPAS (Angwin et al., 2016) are frequently used for bias analyses. The 'AI and Equality Toolbox' (AI and Equality, n.d.) provides a Jupyter Notebook (an interactive document including modifiable code) for exploring biases within the German Credit Dataset (Hofmann, 1994). Other fairness-related datasets can be found in a review by Pessach and Shmueli (2022).

Another pedagogical intervention worthy of attention had students use 'datasheets' (i.e., structured documents that provide contextual information about a dataset) when working on an ML problem (Boyd, 2021). The study found that participants using datasheets identified ethical issues earlier and more often than those without. Similar documents called 'model cards' exist for ML models (Mitchell et al., 2019). Introducing engineers to such tools could potentially help address the dislocated accountability issue related to modularity. A key challenge for educators is that these tools are in their infancy and will likely evolve.

Exercises with models

Having students train an AI model themselves can create interesting conditions for ethical reflection, as suggested by Ko and colleagues: "have students train basic machine learning models, and then reflect on the application and limitation of those models to particular contexts, such as admissions and financial aid decisions" (Ko et al., 2023, chap. 15).

AI models for classification and prediction can be the object of fairness analysis exercises by having students compute and interpret fairness metrics (Pessach & Shmueli, 2022). The 'Human Contexts and Ethics' program of Berkeley (Berkeley CDSS, n.d.) proposes Jupyter Notebooks that include programming tasks using fairness assessment libraries, which can also produce visualizations (Quedado et al., 2022). Thanks to the notebook format, the exercises integrate ethical reflection questions related to the limits of fairness metrics and the contextual nature of fairness. Such approaches could be applied to other ethical issues reviewed previously.

However, as far as we know, evaluation of such methods in terms of impact on student learning is lacking. When reporting on a survey of engineering students that included an AI fairness case study, Orchard and Radke (2023) commented: "students are often able to identify and suggest actions for mitigating the [fairness] issue from a technical standpoint but rarely connect it with broader ethical and societal implications" (p. 15834). Educators and researchers should take this preliminary result as a warning about the limits of addressing ethical concepts such as fairness solely through mathematical and technical lenses. More research is needed to identify how to combine experiential approaches with broader philosophical approaches.

Engineering projects

A central experiential component of engineering education, engineering projects provide evident opportunities to integrate AI ethics. In their 'simplest' form, interventions in projects can build on ethical reflection tools such as questions (Saltz et al., 2019). Assessing students' ethical reflection can be a difficulty for engineering educators, but can be overcome with appropriate pedagogical support (e.g. grading rubrics). Involving ethicists and social scientists in projects is another way to integrate ethics into AI engineering (Tigard et al., 2023), provided that students receive appropriate training and support for interdisciplinary teamwork to ensure a positive experience. Finally, projects involving a human research component can help students develop research ethics skills and methods from the human sciences fields (Williams et al., 2020).

Projects can also provide opportunities for students to practice specific engineering ethics methods applied to AI, a type of intervention we were not able to find in existing publications. We suggest in particular value-sensitive design (Friedman & Hendry, 2019; see Chapter 22); participatory design (Gerdes, 2022; see Chapter 23); ethical risk assessment (Hardebolle et al., 2023); technology assessment (Børsen, 2021); and life-cycle assessment (Ligozat et al., 2022). The main challenge, in this case, is to involve trained specialists of these methods in the design and supervision of the projects and/or to train the teaching teams. This challenge is also an asset – shared with the approaches that we discuss in the next section.

Curriculum-wide interventions

Harvard University implements a curriculum-wide program called "Embedded EthiCS" (Grosz et al., 2019) wherein philosopher-designed ethics modules involving case studies with analytic methodologies and small group discussions are embedded into computer science courses. While some evaluations have been conducted in terms of students' interest and self-efficacy toward ethical issues (Horton et al., 2022), more research is essential to assess the impact of such interventions, particularly for engineers.

Northeastern University chose to embed 'Values Analysis in Design' modules into AI-related courses (Kopec et al., 2023). While the pedagogical methods used are mostly similar to those mentioned above, its specific focus on value analysis builds on prior engineering ethics work with value-sensitive design (Friedman & Hendry, 2019). An evaluation showed significant changes in students' attitudes with respect to values and ethically responsible design (Kopec et al., 2023). While further evaluation is needed, such approaches could also be applied to programs teaching AI to engineers. See Chapter 22 on VSD and Chapter 12 on engineering design for further discussions.

Overall, the contextualization of ethical concerns in practical settings provided by experiential practice-based and embedded approaches seems promising. However, they have their detractors (e.g. Raji et al., 2021), and the evidence is still extremely limited.

AI as a tool for teaching ethics

Applications of AI as a teaching and learning tool are almost as old as the field itself, but the LLM boom has now heightened interest and fears. It would be beyond the scope of this chapter to elaborate on the use of AI for general education, or even for engineering education, and so we focus only on applications to ethics education, a domain which seems under-explored. However, we want to make clear that we by no means consider AI as a silver bullet for this task, not least because it comes with concerning ethical issues that we explore at the end of this section.

AI in case-based learning

Previous work as explored the use of AI for on-the-spot assistance but also for preparatory training in moral decision-making (O'Neill et al., 2022). For instance, AI could be used to provide interactive, personalized, step-by-step guidance in case study analysis. In utilitarian calculus applications, additional information (e.g., background, stakeholder preferences, and probabilities) could be interactively provided to learners. Alternatively, learners could be presented with similar cases to compare since, unlike case law, in ethics precedents are not binding. O'Neill et al. (2022) have, however, flagged critical ethical risks associated with this use – such as unintended influence – and others to which we return later.

Students as critics of AI output

Students' experience with publicly available generative AI could be leveraged for ethics interventions. For instance, students could be asked to create text, images, or videos and analyze the output in terms of the kind of values or biases they present (e.g., political bias, see Narayanan, 2023) or to identify instances of 'plausible non-sense' in AI chatbot outputs (Hardebolle & Ramachandran, 2023) and reflect on how much such systems should be trusted. While we found academic work doing this type of analysis (e.g., Srinivasan & Uchino, 2021), we were not able to find studies on educational interventions. One challenge is that methodologies for performing such evaluations rigorously can be quite complex. It is worth highlighting that some studies found that students may be reluctant to use generative AI tools even when encouraged to do so (Prasad et al., 2023).

For classification or prediction models, students could be guided to use one such model to make a decision and then reflect on how they made the decision, particularly in terms of their own cognitive biases. Such an activity could provide an introduction to the challenges of AI-assisted decision-making, particularly the issue of automation bias (Suresh et al., 2020). The effectiveness and challenges of such interventions remain to explore.

Students as 'subjects' of AI processing

Prior research has explored activities where students have worked with data on themselves to increase learning engagement. Shapiro et al. (2020) showed how such activities can support critical reflection and help students develop an ethics of care. Although they are preliminary, these results seem promising as "students were confronted with the idea that they are the 'other' within systems that use and may exploit personal data and as a result, began to consider what care they desire or demand from these systems" (Shapiro et al., 2020, p. 9).

A similar 'making it personal' approach has been explored in AI-type tasks reported by Register and Ko (2020). Students implemented a model that predicts a student's grade based on self-reported measures of interest in courses and input their own data. Although the intervention was limited to very simple models, students seemed to pay more attention to the teaching material and appeared better able to explain underlying ML mechanisms. We see potential interest in these methods for students to empathize with end-users, realize what AI-assisted decision-making means in practice, and get a better understanding of transparency issues. Still, these hypotheses would need to be tested.

Even AI tools that may be generally considered 'harmful' or 'unethical' could be used as pedagogical instruments to explain the consequences of their irresponsible use on unwitting stakeholders. For instance, Ramachandran et al. (2023) examined the effect of using deepfakes as a pedagogical tool to foster students' empathy towards victims of this technology (as in the case of

non-consensual deepfakes, pornographic ones in particular). This topic appeals to the students' sense of responsibility as potential creators of AI tools. Engaging in such discussions prompts students to confront similar ethical quandaries they may encounter as professionals in the future, enhancing their moral sensitivity, motivation, and reasoning.

Another way of making it personal is to have an AI assess students' productions (such as essays) and guide students to reflect on the process and its results afterward. While evaluations of the potential of AI for this type of use exist, we did not find interventions that make use of it for teaching ethics. Such an intervention could provide opportunities for discussing the ethics of automated evaluation, trustworthiness, transparency, and empowerment questions, as well as the role of emotions in ethics. However, instructors should exercise caution and assess both the ethics and legality of this type of setup in their own context.

Overall, the balance between benefits and risks of all the interventions mentioned in this section should be carefully evaluated, a point we address in the next section.

Ethics of using AI for EEE

In this section, we examine the ethical risks associated with the use of AI in EEE by successively adopting the point of view of the five 'ethical lenses' of the 'Digital Ethics Canvas' (Hardebolle et al., 2023), a methodological tool designed for teaching ethical risk assessment to engineers.

Sustainability

Encouraging AI usage in universities raises systemic environmental risks as the rising energy consumption and resultant carbon footprint from server operations per user is immediately multiplied by large numbers of students. As we have seen, the environmental impact of generative AI is much higher than most other types of software or digital tools (Luccioni et al., 2023). Instructors ought to evaluate the necessity of using AI systems for specific educational tasks, and consider alternatives that have a lower impact. More generally, environmental impact and labor practices should be treated as essential criteria when selecting an AI system.

Privacy

Student privacy and data security is of prime importance in educational contexts. The collection of student data for AI use in ethics education is a real risk since it may include sensitive information about student values and morality (O'Neill et al., 2022). The potential re-use of student data for AI training is also of concern since training data can be retrieved from models (Carlini et al., 2023). While European institutions are particularly attentive to institutional use with the General Data Protection Regulation (GDPR), it is imperative to sensitize students to data consent and its consequences, especially with US-hosted tools such as ChatGPT.

Fairness

Students should not incidentally be subjected to unfair treatment or outcomes while using AI for ethics education. Two aspects to consider for fair treatment are access and accessibility. Although free accounts can facilitate access, they often lead to problematic differences in privacy treatment. Accessibility considerations (e.g., interface, language) are often not considered in software interfaces, and AI is no exception. Regarding outcomes, although demonstrating the biases in AI-generated output (Abid et al., 2021) can be helpful as an educational exercise, instructors

should not underestimate the emotional response or even trauma that exposure to biased information can generate and must take appropriate measures.

Non-maleficence

Significant attention has been drawn to the potential adverse effects of generative AI on human learning, even though some argue that this issue dates back to the invention of writing (see Plato, Phaedrus 14, pp. 274–275). The impact on human skills generally requires more research. An open question regards whether AI harms the learning assessment process: on one hand, it interferes with students' writing; on the other hand, it can be used by instructors to ease the tedious process of analyzing textual productions (which comes with other risks, as discussed earlier). In addition, we should not lose sight of the harms that arise at a more macro/global level, among which we can cite content stolen from authors and artists – and information pollution on a large scale.

Empowerment

The 'plausible nonsense' (also called 'hallucinations,' see Huang et al., 2023 for a review) unpredictably generated by LLMs might offer intriguing exercises for practicing critical thinking. However, aggravated by the lack of information provided to users on the unreliability of the output, it remains problematic in numerous scenarios (e.g., searching for information). When available, AI tools that provide ways for users to evaluate output quality are generally preferable, particularly in educational settings. In addition to dis-empowerment risks relating to the black-box nature of AI and the associated "inescapability of outside influence" (O'Neill et al., 2022, p. 9), some interface designs can also increase the human tendency to anthropomorphize these systems, which can lead to serious consequences (manipulation, in particular, emotional manipulation and dependency), particularly for vulnerable groups.

With our review of the ethical risks above (which is not exhaustive), we hope that we have illustrated how critical reflective practice can be applied to the case(s) of using AI tools in ethics education. Beyond AI, our use of digital tools in education should be driven by our values – an exercise that is challenged by the pressure of productivity and the strong push from tool vendors.

Conclusions

This chapter has grappled with a unique set of challenges and opportunities. Although the topics of AI ethics and AI in education are rich and constantly evolving, pedagogical methods are still nascent, particularly within the context of engineering education. We navigate the inherent complexities of this field by adopting an interdisciplinary view that balances our varying opinions. However, we are simultaneously unwavering in our commitment to addressing the broad spectrum of ethical issues that arise when AI is used in education. One of the limitations of this chapter, and a challenge for EEE practitioners and researchers, is the temporality of our conclusions – AI and AI ethics evolve at lightning speed as new technologies, policies, and ethical dilemmas emerge.

Acknowledgments

Special thanks to Jules Ronné for the biomechanical engineering examples.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*(6), 461–463. doi: 10.1038/s42256-021-00359-2

AI and Equality. (n.d.). *A human rights toolbox.* Retrieved 2024-01-08, from https:// aiequalitytoolbox.com/

AI-HLEG. (2019). *Ethics guidelines for trustworthy ai* (Tech. Rep.). European Commission, B-1049 Brussels: European Commission.

Alam, A. (2023). Developing a curriculum for ethical and responsible AI: A university course on safety, fairness, privacy, and ethics to prepare next generation of AI professionals. In G. Rajakumar, K.-L. Du, & A. Rocha (Eds.), *Intelligent communication technologies and virtual mobile networks* (pp. 879–894). Springer Nature. doi: 10.1007/978-981-99-1767-9_64

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). Machine bias. *ProPublica*. Retrieved 2023-12-20, from https://www.propublica.org/article/ machine-bias-risk-assessments-in-criminal-sentencing

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* (pp. 610–623). ACM. doi: 10.1145/3442188.3445922

Berkeley CDSS. (n.d.). *Human contexts and ethics.* Retrieved August 1, 2024, from https://data.berkeley.edu/human-contexts-and-ethics

Børsen, T. (2021). *Using technology assessment in technical study programs as a means to foster ethical reflections on the societal effects of technologies and engineering solutions*. Proceedings of the SEFI 49th Annual Conference. Berlin, Germany*,* 685–694.

Boyd, K. L. (2021). Datasheets for datasets help ML engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 438:1–438:27. doi:10.1145/3479582

Burton, E., Goldsmith, J., & Mattei, N. (2018). How to teach computer ethics through science fiction. *Communications of the ACM* , *61*(8), 54–64. doi:10.1145/3154485

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). *Extracting training data from diffusion models.* arXiv. doi: 10.48550/ arXiv.2301.13188

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, *41*(3), 1–39.

Coeckelbergh, M. (2022). *The political philosophy of ai: an introduction*. John Wiley & Sons.

Council of the European Union. (2023, November). *Artificial* intelligence act—preparation *for the trilogue. interinstitutional file 2021/0106(COD), Document 15688/23.* Council of the European Union.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Cummings, M. L. (2023, July). *What self-driving cars tell us about ai risks - IEEE spectrum.* Retrieved October 24, 2023, from https://spectrum.ieee.org/ self-driving-cars-2662494269

Dubber, M. D., Pasquale, F., & Das, S. (2020). *The Oxford handbook of ethics of AI*. Oxford Handbooks.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2021). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. In L. Floridi, (Eds.), *Ethics, Governance, and Policies in Artificial Intelligence*, vol 144, Springer, Cham.

Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination*. The MIT Press.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437.

Garrett, N., Beard, N., & Fiesler, C. (2020, February). More than "if time allows": The role of ethics in AI education. In *Proceedings of the AAAI/ACM conference on ai, ethics, and society* (pp. 272–278). ACM. doi: 10.1145/ 3375627.3375868

Gerdes, A. (2022). A participatory data-centric approach to AI ethics by design. *Applied Artificial Intelligence*, *36*(1), 2009222. doi: 10.1080/08839514.2021.2009222

Griffin, T. A., Green, B. P., & Welie, J. V. M. (2023, January). The ethical agency of AI developers. *AI and Ethics*. doi: 10.1007/s43681-022-00256-3

Grosz, B. J., Grant, D. G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded EthiCS: Integrating ethics across CS education. *Communications of the ACM* , *62*(8), 54–61. doi: 10.1145/3330794

Grubb, M. J. (1990). Communication Energy efficiency and economic fallacies. *Energy Policy*, *18*(8), 783–785. doi: 10.1016/0301-4215(90)90031-X

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

Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics*, *2*(4), 851–867. doi: 10.1007/s43681-021-00122-8

Hardebolle, C., Kovacs, H., Simkova, E., Pinazza, A., Di Vincenzo, M. C., Jermann, P., Tormey, R., & Dehler Zufferey, J. (2022). *A game-based approach to develop engineering students' awareness about artificial intelligence ethical challenges*. Proceedings of the 50th SEFI Annual Conference. Barcelona, Spain, 2276–2281. doi: 10.5821/conference-9788412322262.1283

Hardebolle, C., Macko, V., Ramachandran, V., Holzer, A., & Jermann, P. (2023). *The digital ethics canvas: A guide for ethical risk assessment and mitigation in the digital domain*. Proceedings of the 51st SEFI Annual Conference 2023. Dublin, Ireland.

Hardebolle, C., & Ramachandran, V. (2023, October). *Plausible nonsense and carbon footprint: Micro and macro Ethics of generative AI in the classroom.* Retrieved October 31, 2023, from https://www.sefi.be/2023/10/14/plausible-nonsense-and-carbon-footprint-micro-and-macro-ethics-of-generative-ai-in-the-classroom/

Héder, M. (2020). A criticism of AI ethics guidelines. *Információs Társadalom*, *20*(4), 57–73. doi: 10.22503/inftars.XX.2020.4.5

Héder, M. (2021). AI and the resurrection of technological determinism. *Információs Társadalom*, *21*(2), 119–130. doi: 10.22503/inftars.XXI.2021.2.8

Héder, M. (2023a). The epistemic opacity of autonomous systems and the ethical consequences. *AI & SOCIETY*, *38*, 1819–1827. doi: 10.1007/s00146-020-01024-9

Héder, M. (2023b). Explainable AI: A brief history of the concept. *ERCIM NEWS, 134*, 9–10.

HHS OCAIO. (2023). *Artificial intelligence use cases – FY2022.* Retrieved December 20, 2023, from https://www.hhs.gov/sites/default/files/hhs-artificial-intelligence -select-use-cases.pdf

Hingle, A., & Johri, A. (2023). *Recognizing principles of AI Ethics through a role-play case study on agriculture*. ASEE Annual Conference and Exposition, Conference Proceedings.

Hofmann, H. (1994). *German credit data (statlog)).* UCI Machine Learning Repository. doi: 10.24432/C5NC77

Horton, D., McIlraith, S. A., Wang, N., Majedi, M., McClure, E., & Wald, B. (2022). Embedding Ethics in computer science courses: Does it work? In *Proceedings of the 53rd ACM* technical symposium *on* computer science education (Vol. 1, pp. 481–487). ACM. doi: 10.1145/3478431.3499407

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions* (arXiv:2311.05232). arXiv. https://doi.org/10.48550/arXiv.2311.05232

Ienca, M. (2023). On artificial intelligence and manipulation. *Topoi*. doi: 10.1007/ s11245-023-09940-3

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

Ko, A. J., Beitlers, A., Wortzman, B., Davidson, M., Oleson, A., Kirdani-Ryan, M., Druga, S., & Everson, J. (2023). *Critically conscious computing: Methods for secondary education*. Retrieved October 23, 2023, from https://criticallyconsciouscomputing.org

Kopec, M., Magnani, M., Ricks, V., Torosyan, R., Basl, J., Miklaucic, N., Muzny, F., Sandler, R., Wilson, C., Wisniewski-Jensen, A., Lundgren, C., Baylon, R., Mills, K., & Wells, M. (2023). The effectiveness of embedded values analysis modules in computer science education: An empirical study. *Big Data & Society*, *10*(1), 20539517231176230. doi: 10.1177/20539517231176230

LaChance, J., Thong, W., Nagpal, S., & Xiang, A. (2023). *A case study in fairness evaluation: Current limitations and challenges for human pose estimation*. AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI.

Larosa, F., Hoyas, S., García-Martínez, J., Conejero, J. A., Fuso Nerini, F., & Vinuesa, R. (2023). Halting generative AI advancements may slow down progress in climate research. *Nature Climate Change*, *13*(6), 497–499. doi: 10.1038/s41558-023-01686-5

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). *Making AI Less "thirsty": Uncovering and addressing the secret water footprint of AI models.* arXiv. doi: 10.48550/ arXiv.2304.03271

Ligozat, A.-L., Lefevre, J., Bugeau, A., & Combaz, J. (2022). Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. *Sustainability*, *14*(9), 5172. doi: 10.3390/su14095172

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Luccioni, A. S., Jernite, Y., & Strubell, E. (2023). *Power hungry processing: Watts driving the cost of AI deployment?* arXiv. doi: 10.48550/arXiv.2311.16863

Madiega, T. (2021). Artificial intelligence act. *European Parliament: European Parliamentary Research Service*.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*, 175–183.

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI ethics: Relational autonomy as a means to counter AI harms. *Topoi*, *42*(3), 867–880. doi: 10.1007/s11245-022-09874-2

MIT. (n.d.). *Moral machine.* Retrieved January 8, 2024, from http://moralmachine.mit.edu

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220–229). doi: 10.1145/3287560.3287596

Narayanan, A. (2023, March). *Does ChatGPT have a liberal bias?* Retrieved November 2, 2023, from https://www.aisnakeoil.com/p/does-chatgpt-have-a-liberal-bias

Orchard, A., & Radke, D. (2023). An analysis of engineering students' responses to an AI Ethics scenario. In B. Williams, Y. Chen, & J. Neville, (Eds.), *Proceedings of the 37th AAAI conference on artificial intelligence, AAAI 2023* (Vol. 37, pp. 15834–15842).

O'Neill, E., Klincewicz, M., & Kemmer, M. (2022). Ethical issues with artificial ethics assistants. In C. Véliz (Ed.), *The Oxford handbook of digital ethics*. Oxford University Press. doi: 10.1093/oxfordhb/9780198857815.013.17

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). *Carbon emissions and large neural network training.* arXiv. doi: 10.48550/arXiv.2104.10350

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys*, *55*(3), 51:1–51:44. doi: 10.1145/3494672

Petrozzino, C. (2021). Who pays for ethical debt in AI? *AI and Ethics*, *1*(3), 205–208. doi: 10.1007/s43681-020-00030-3

Prasad, S., Greenman, B., Nelson, T., & Krishnamurthi, S. (2023). Generating programs trivially: Student use of large language models. *Proceedings of the ACM Conference on Global Computing Education*, *1* . doi: https://doi.org/10.1145/ 3576882.3617921

Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence*, *4*(2), 99–101.

Quedado, J., Zolyomi, A., & Mashhadi, A. (2022). A case study of integrating fairness visualization tools in machine learning education. In *Extended abstracts of the 2022 CHI conference on human factors in computing systems.* ACM. doi: 10.1145/3491101.3503568

Raji, I. D., Kumar, I. E., Horowitz, A., & Selbst, A. (2022). The fallacy of AI functionality. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 959–972). ACM. doi: 10.1145/3531146.3533158

Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2021). You can't sit with us: Exclusionary pedagogy in AI ethics education. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 515–525). New York, NY, USA: ACM. doi: 10.1145/3442188.3445914

Ramachandran, V., Hardebolle, C., Kotluk, N., Ebrahimi, T., Riedl, R., & Jermann, P. (2023). *A multimodal measurement of the impact of deepfakes on the ethical reasoning and affective reactions of students*. Proceedings of the 51st SEFI Annual Conference 2023. Dublin, Ireland.

Register, Y., & Ko, A. J. (2020). Learning machine learning with personal data helps stakeholders ground advocacy arguments in model mechanics. In *Proceedings of the 2020 ACM conference on international computing education research* (pp. 67–78). ACM. doi: 10.1145/3372782.3406252

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence a modern approach*. Prentice hall.

Sætra, H. S., & Danaher, J. (2023). Resolving the battle of short- vs. long-term AI risks. *AI and Ethics*. doi: 10.1007/s43681-023-00336-y

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N, & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education*, *19*(4), 32:1–32:26. doi: 10.1145/3341164

Sanderson, C., Douglas, D., & Lu, Q. (2023). Implementing responsible AI: Tensions and trade-offs between ethics aspects. In *2023 International joint conference on neural networks (IJCNN)* (pp. 1–7). doi: 10.1109/IJCNN54540.2023.10191274

Sattlegger, A., van den Hoven, J., & Bharosa, N. (2022). Designing for Responsibility. In *DG.O 2022: The 23rd annual international conference on digital government research* (pp. 214–225). ACM. doi: 10.1145/3543434.3543581

Scharowski, N. (2020). *Transparency and trust in AI* (Unpublished doctoral dissertation). Institute of Psychology.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278.

Shapiro, B. R., Meng, A., O'Donnell, C., Lou, C., Zhao, E., Dankwa, B., & Hostetler, A. (2020). Re-shape: A method to teach data ethics for data science education. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13). ACM. doi: 10.1145/3313831.3376251

Srinivasan, R., & Uchino, K. (2021). Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 41–51).

Stark, L., & Hoey, J. (2021). The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 782–793). Association for Computing Machinery. https://dl.acm.org/doi/proceedings/10.1145/3442188

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization* (pp. 1–9). ACM. doi: 10.1145/3465416.3483305

Suresh, H., Lao, N., & Liccardi, I. (2020). Misplaced trust: Measuring the interference of machine learning in human decision-making. In *12th ACM conference on web science* (pp. 315–324). ACM. doi: 10.1145/3394231.3397922

Tigard, D. W., Braun, M., Breuer, S., Ritt, K., Fiske, A., McLennan, S., & Buyx, A. (2023). Toward best practices in embedded ethics: Suggestions for interdisciplinary technology development. *Robotics and Autonomous Systems*, *167*, 104467. doi: 10.1016/j.robot.2023.104467

Tuovinen, L., & Rohunen, A. (2021). Teaching AI ethics to engineering students: Reflections on syllabus design and teaching methods. In *Proceedings of AAAI'23/IAAI'23/EAAI'23* (pp. 15834–15842). doi: 10.1609/aaai.v37i13.26880

Widder, D. G., & Nafus, D. (2022). *Dislocated accountabilities in the AI supply chain: Modularity and developers' notions of responsibility.* arXiv. doi: 10.48550/arXiv.2209.09780

Williams, T., Zhu, Q., & Grollman, D. (2020, April). An experimental ethics approach to robot ethics education. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), 13428–13435. doi: 10.1609/aaai.v34i09.7067

Winfield, A. F., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R. I., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M. A., Wortham, R. H., & Watson, E. (2021). IEEE P7001: A proposed standard on transparency. *Frontiers in Robotics and AI*, *8*, 665729.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com