An Analysis of Engineering Students' Responses to an AI Ethics Scenario

Alexi Orchard[1], **David Radke**[2]

[1] Department of English Language and Literature, University of Waterloo, Waterloo, Canada
[2] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada
{alexi.orchard, dtradke } @uwaterloo.ca

Abstract

In light of significant issues in the technology industry, such as algorithms that worsen racial biases, the spread of online misinformation, and the expansion of mass surveillance, it is increasingly important to teach the ethics and sociotechnical implications of developing and using artificial intelligence (AI). Using 53 survey responses from engineering undergraduates, this paper measures students' abilities to identify, mitigate, and reflect on a hypothetical AI ethics scenario. We engage with prior research on pedagogical approaches to and considerations for teaching AI ethics and highlight some of the obstacles that engineering undergraduate students experience in learning and applying AI ethics concepts.

Introduction

With the rise of significant ethical problems in the technology industry, such as algorithms that worsen racial biases, the spread of online misinformation, and the expansion of mass surveillance (Noble 2018; Vraga, Tully, and Bode 2020; Zuboff et al. 2019), it is important to engage topics relating to ethics, fairness, psychological impact, social and environmental justice, and equity, diversity, and inclusion in artificial intelligence (AI) curriculum. To accomplish this, we must examine how educators and institutions prepare future graduates to identify and address ethical problems.

Scholars agree that the historically isolated ethics ecosystem of engineering and tech-related disciplines, combined with ongoing social injustices exacerbated by irresponsible tech design in industry, points to a need for ethics to have a more rigorous presence in the engineering curriculum (Raji, Scheuerman, and Amironesei 2021; Nasir et al. 2021; Antoniou 2021). Many institutions have introduced AI ethics courses in recent years; since 2019, educators and researchers have crowd-sourced and catalogued upwards of 400 AI and/or tech ethics (including content on machine learning and AI algorithms) syllabi from North American institutions (Raji, Scheuerman, and Amironesei 2021; Fiesler, Garrett, and Beard 2020; Nasir et al. 2021). Educators have proposed and implemented individual modules, workshops, and heuristics for teaching AI ethics in tandem with their standard coursework (Cohen et al. 2021; Saltz et al. 2019; Furey and Martin 2019). Some of these curricular interventions received student feedback that the content was relevant and interesting, and they would like to learn more throughout their program (Grosz et al. 2019; Cohen et al. 2021). To this end, the amount of resources and tools for teaching AI ethics has grown steadily in the last decade; meanwhile, educators are still in the early stages of evaluating these pedagogical strategies and measuring student perceptions, retention, and interest in the topic.

Our study enters this conversation by analyzing 53 survey responses from engineering undergraduate students on an AI ethics problem wherein a facial recognition model fails to accurately identify dark-skinned faces. In qualitative short answer responses, participants were prompted to explain how they would respond and what they think are the most and least important considerations in the given scenario. Our results indicate that students are often able to identify and suggest actions for mitigating the issue from a technical standpoint but rarely connect it with broader ethical and societal implications.

This paper reviews various pedagogical approaches to teaching AI ethics, and relates these methods with our analysis of student knowledge, attitudes toward, and considerations of AI ethics from the survey responses. Our survey also asked participants to describe any obstacles they have experienced in learning and applying AI ethics concepts in their curricular and work experiences. We propose that teaching AI ethics with a sociotechnical pedagogical model would complement students' existing technical skills and enhance their ability to ask critical questions in mitigating complex ethical issues. By observing and reflecting on student responses in conjunction with current ethics pedagogy and research, we anticipate that this paper will inform future pedagogical design and strategies for integrating AI ethics into the engineering curriculum.

Background

Engineering and Computing Ethics

Ethics was included in the American Board of Engineering and Technology (ABET) accreditation requirements in 2000; the Canadian Engineering Accreditation Board (CEAB) added a similar attribute nearly a decade later (Roncin 2013). "Engineering ethics," in accreditation

contexts, describes a rather narrow understanding of what ethics entails; for example, the code states that engineers are required to "hold paramount the safety, health, and welfare of the public," but there is limited clarification on who and what "the public" includes (Hipp 2007).

The what and how of integrating ethics into engineering, computer science, and other technology-oriented programs has been discussed for many years: prior research debates the merits of technical courses with ethics added in, non-technical courses characterized by philosophical or moral debate, and other approaches that engage humanities and social science perspectives (e.g., Science and Technology Studies, Critical Data Studies) (Hess and Fore 2018; Herkert 2000). Institutions often utilize a combination of these approaches depending on their resources (Walczak et al. 2010). Engineering students are typically exposed to ethics through professional codes of conduct and case studies (Walczak et al. 2010; McGinn 2018). Engineering graduates who become licensed by their professional association are held accountable to a Code of Ethics. Typical case studies include the 1907 Quebec bridge collapse, the Space Shuttle Challenger disaster, and Bhopal Plant disaster.

The emphasis on ethics within computing and engineering programs has increased in recent years (Green 2021). One 2021 syllabi review compiled 254 AI ethics courses at 132 North American universities (Raji, Scheuerman, and Amironesei 2021). The AI ethics research community, too, has grown with conferences such as the ACM Conference on Fairness, Accountability, and Transparency (FAccT) and the AAAI/ACM Conference on AI, Ethics, and Society (AIES).

While this paper focuses on engineering students and courses, we recognize there is considerable overlap with computer science in regard to notions of computing and tech ethics. However, we emphasize the distinction between *engineering* ethics and computing and tech because of the professional associations and codes that engineering graduates frequently belong and adhere to. The nearest equivalent in computer science is the codes of ethics presented by professional computing organizations such as the IEEE and ACM but their codes are not enforceable, nor are they required to be taught (Mittelstadt 2019).

Computing and tech ethics are illustrated in ethical and responsible design manifestos, declarations, and principles produced by academia, industry, and government entities (Mulvenna, Boger, and Bond 2017; Mozilla 2007; The Future of Life 2017). These documents advocate for principles such as beneficence, autonomy, fairness, sustainability, and privacy (Morley et al. 2021), and can also be used for teaching engineering and tech ethics. According to the Global Inventory of AI Ethics Guidelines, managed by Algorithm Watch, there are now more than 160 documents in existence (last updated in 2020) (Chiusi 2020). Though these documents may be well-intended, a major criticism is that they are often too abstract to be actionable and may be used for corporate "ethics-washing" more so than addressing real issues when innovating (Green 2021). The broad framing of AI and tech ethics here makes it difficult for students and instructors, and those in industry, to learn and apply ethical and responsible design in their respective contexts.

Obstacles and Considerations in Ethics Education

Instructors Though there are a wealth of pedagogical resources and methods for teaching ethics, there are a number of obstacles for instructors. Engineering instructors, in many cases, do not have a background and are therefore untrained in or uncomfortable broaching ethical topics (Johnson 1994; Walczak et al. 2010). In many cases, instructors are not incentivized and supported by their institution (and, in turn, the accreditation boards) to incorporate more ethics into an already tightly packed curriculum (Walczak et al. 2010).

When taught by non-technical instructors, often from a humanities or social science background, students benefit from another perspective on the ethical implications of technology; however, one drawback may be they are less likely to hear the importance of ethics reinforced by someone in their primary discipline and have more difficulty tying ethical concepts to their technical material. To this point, a recent study that integrated equity, diversity, and inclusion (EDI) modules into their engineering courses observed that students, particularly those from marginalized communities, appreciated learning the EDI content from their home department and instructors (d'Entremont et al. 2022).

Multidisciplinary, collaborative ethics teaching (e.g., cross-faculty team-teaching) has shown promise in exposing students to complex sociotechnical discussions (Hoople and Choi-Fitzpatrick 2017) but these efforts meet logistical challenges related to scheduling, accreditation, and institutional policy (Walczak et al. 2010).

Content Krakowski et al. observes that, much like engineering and computing ethics content generally, AI ethics training is done separately from technical coursework, resulting in a decontextualization of ethical concerns from real-world consequences and a de-emphasis of ethical issues in comparison with technical content (Krakowski et al. 2022). Integrating ethics across the curriculum, through additional readings, modules, and non-technical instructor expertise, has been been successful previously (Grosz et al. 2019); however, other studies observe that individual efforts (without consistent framing and support of ethics' relevance and importance from more than one source) could suggest that ethics content is *supplemental* rather than *fundamental* to students' training (Cech 2014; Garrett, Beard, and Fiesler 2020). Achieving this kind of cross-curricular integration relies on the coordination of instructors, faculty, and administrative figures to support such an initiative.

In a review of AI ethics syllabus design and teaching methods, Tuovinen and Rohunen emphasized how ethical questions and topics must strike a balance with technical content to allow for students to see the significance and applicability of the ethical content (Tuovinen and Rohunen 2021). Some aspects of ethics can be presented as facts, such as the safety record and current legislation governing a technology's development and use, but ethics content on a deeper level is inherently subject to ambiguity and debate (Tuovinen and Rohunen 2021). As such, instructors must consider the delivery of this content, in addition to the content itself, to best facilitate the learning process for students.

Students Though all the above mentioned obstacles influence students, we recognize that experiences outside of the classroom also shape students' engagement with ethics in their education and career. Cooperative education (co-op), referred to as internships or work-integrated study at some institutions, is a driving motivation for many students – particularly engineering and computer science students on the verge of becoming high income earners in the tech industry. Truax et al. observed that while coursework aided students in recognizing the importance of engineering ethics, they seldom saw the opportunity to apply the concepts in other projects or on co-op jobs (Truax, Orchard, and Love 2021), a theme also discussed in our study's results.

Though the consideration of ethics is an essential part of engineering, there are still many barriers hindering the impact of ethics in the curriculum.

Teaching AI Ethics

The current curriculum prepares future graduates to be competent, resourceful problem solvers in their technical fields; however, as we have noted, many obstacles make it challenging for ethics and responsible design to be integrated into tech curriculum and the engineer's workflow. In this section, we outline some of the pedagogical strategies and frameworks used in AI ethics curriculum and review social and technical perspectives on bias in facial recognition technology (FRT), the focus of the scenario used in our survey.

Frameworks for Teaching AI Ethics

In a typical AI ethics course, instructors may discuss relevant ethical principles (e.g. found in a professional code or a tech ethics manifesto) and their interpretation in practice, broader societal implications (e.g. implications of FRT for mass surveillance), or downstream developments and consequences of future technology (e.g. artificial general intelligence) (Tuovinen and Rohunen 2021; Fiesler, Garrett, and Beard 2020). Themes found in AI ethics courses include data ethics, privacy, AI literacy, legislation, and accountability and transparency, among others; these concepts are often disseminated through articles, stories, film, and discussion-based exercises and assignments (Tuovinen and Rohunen 2021).

Educators have also explored the development of new or adoption of existing frameworks for identifying ethical issues. For instance, Saltz et al.'s framework includes a list of questions (i.e., "How might an individuals' privacy and anonymity be impinged via aggregation and linking of the data?") for students to address ethical issues in their machine learning (ML) projects. This framework was designed to be directly inserted into their existing technical ML modules on logistic regression, random forest classification, and various kinds of deep learning models. Their results show that students can easily apply an explicit set of questions to a ML project, focusing specifically on "ethical issues that are actionable by members of an ML project team, and not those that are societal in nature" (Saltz et al. 2019).

Krakowski et al. also presented a framework that prompts students to (1) determine whether AI is an appropriate tool for the defined task, (2) question the data being used, (3) consider the affordances and limitations of the AI system's design, and (4) consider how the AI system's output will impact real world systems (Krakowski et al. 2022). They measured this approach with pre/post surveys, interviews, and focus groups that prompted students to identify areas of concerns about a proposed AI system or its deployment. Their results indicate an increased level of sophistication in the students' abilities to integrate ethics concerns with technical features of an AI system (Krakowski et al. 2022). Though this framework was initially tested with high school students, who do not have the same technical background as undergraduate engineers, it appears to be a promising entry point that could be integrated and expanded into more advanced curricula.

Krakowski et al.'s framework has a broader scope of what constitutes an ethical issue than Saltz et al.'s, while Saltz et al. provides a direct entry into technical ML curricula; nevertheless, aspects of each approach could be derived to suit the focus of different curricular designs. The Discussion section will put these frameworks into conversation with the FRT scenario used in our study.

Case Study: Facial Recognition Technology

Facial recognition technology is a broad term to describe biometric software that can be used for facial detection, analysis, and verification or identification of a human subject. For example, it can be used to determine physical or demographic traits like age, gender, or facial expression. In this section, we provide a brief overview of technical and ethical considerations of algorithmic bias in FRT. This overview will provide context for how an AI ethics framework could be applied to the FRT scenario used in our study.

Bias in Deep Learning Models Facial recognition models are typically taught within the scope of supervised deep learning, where the model fits to a given dataset and can be used to extrapolate decisions about new data from a similar distribution. This is done by training the model on the dataset, repeatedly tuning the model's numerical parameters to obtain higher accuracy.

One way a model can become biased is by learning to perform a task significantly better on a subset of the data distribution. The reason why this happens can be difficult to understand; however, one cause may be an unbalanced dataset (i.e., more examples of a particular group). Attempting to mitigate bias with unbalanced data is an active area of research within AI, and some solutions can be divided into two broad categories: modifying the original dataset or modifying the learning process.

One method to mitigate the problem of learning bias from unbalanced data is to obtain more data of under-represented classes; however, this may not always be possible when data is sparse. One could then enrich under-represented classes using data augmentation, the process of generating artificial data from a smaller set of existing real samples (Taylor and Nitschke 2018). A more advanced technique to generate data utilizes Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) to learn the dataset's underlying distribution

and generate new realistic samples for a smaller class (Perez and Wang 2017; Cubuk et al. 2019).

Other methods directly impact how models learn from potentially unbalanced datasets by limiting the amount that parameters fit to over-represented groups (Kenfack et al. 2022). Another approach stems from research on techniques to help train deep learning models with less data (Srivastava et al. 2014; Ioffe and Szegedy 2015; Weiss, Khoshgoftaar, and Wang 2016; Erhan et al. 2010), meaning the model could then learn from a smaller, but balanced dataset.

In AI education, students typically learn several of these techniques to overcome challenges when training models.

Ethical Implications of Facial Recognition Technology Learning to identify and mitigate bias and unbalanced data is a common outcome in a technical AI course. However, being able to recognize potential consequences of such an issue requires more analysis of the social and ethical context of the technology itself. Many applications of FRT, whether on social media, at a border station, or medical scanning (Hare 2022), are susceptible to misuse and can result in discrimination against marginalized groups (Benjamin 2019).

Some of the main ethical tensions of FRT include privacy and representation, intersectionality and group-based fairness, and transparency and overexposure (Raji et al. 2020a). For example, though we may aim to account for diverse representation in a dataset, this can present privacy risks, issues of consent, and perpetuate marginalization of certain populations (Hamidi, Scheuerman, and Branham 2018; Hoffmann 2019; Mozur, Paul 2019). Furthermore, it is difficult to achieve equal representation among underrepresented subgroups and, in the case that some balance is achieved, it may come at the cost of elevating other risks (Raji et al. 2020a).

Algorithmic audits, one method for detecting issues of fairness and accountability, are "assessments of the algorithm's negative impact on the rights and interests of stakeholders, with a corresponding identification of situations and/or features of the algorithm that give rise to these negative impacts" (Brown, Davidovic, and Hasan 2021). There has been growing attention to algorithmic auditing in AI ethics research in recent years (Mittelstadt 2016; Raji et al. 2020b; Vecchione, Levy, and Barocas 2021); integrating it into AI ethics curricula is an area for future research.

In summary, building, modifying, or expanding a dataset to mitigate bias demands critical questioning about its purpose and potential implications that need to be considered alongside deep learning techniques.

Methods

Survey Instrument

This study utilizes a mixed-methods qualitative and quantitative survey with sections dedicated to three hypothetical scenarios and personal definitions of ethics and responsible design. This study is one part of a larger project investigating engineering and tech ethics more broadly, so other scenarios in the survey are placed in different engineering contexts; in this paper, we only analyze the responses from the AI ethics scenario. Additionally, this survey is not in response to any specific pedagogical intervention but instead serves as a baseline measure for engineering student knowledge of an AI ethics problem. We discuss our plans for interpreting the baseline results in the future research section. Students are prompted to describe how they would respond to this scenario:

> **Scenario:** For a final project, you acquire a dataset to build a facial recognition model to predict a subject's age. You train a model, and your initial results achieve 95 percent accuracy on the testing set. When you dig into the 5 percent incorrect samples, you realize that the accuracy is very high for lighter skinned individuals but very low for darker skinned individuals. Your project's intended population is mostly accounted for within the test set. You know that your project will surely receive a high grade if you report your initial results.

Prior research on the use of hypothetical scenarios finds they are useful for emulating potentially sensitive ethical topics (Aguinis and Bradley 2014). When teaching ethics, scenarios should be relevant and familiar yet generalizable (Hishiyama and Shao 2022; Weber 1992).

In this scenario, we sought to find a balance between the relatable context of coursework and a real-world concern by proceeding with the assumption that the issues associated with facial recognition technology are well-known and documented in both the tech sector and popular news outlets (Van Noorden 2020). This was done to avoid drawing on the paradigmatic case studies of catastrophic design failures often used in ethics training (Hipp 2007), while still gesturing towards authentic considerations in AI. Other details in the scenario, such as the 5 percent error, the intended population, and the project grade, represented factors that could influence the participants' actions, much like design constraints, pressure to meet deadlines, and workplace dynamics might also factor into a real world context.

The short answer section asks participants to describe how they would respond and what they think are the most and least important considerations in the given scenario. They were also asked to describe any obstacles they have previously experienced in applying ethical concepts in their coursework, projects, and co-op workplaces. Short answer responses are measured by the quantity and type of considerations raised and measured sample-wide thematically.

Participants

This study targeted students across various Engineering programs during the Winter 2022 term. They were invited to participate in connection with our larger research project; none of the instructors who volunteered to advertise this survey to their courses are part of the research team nor do they necessarily teach any ethics content themselves.

The survey was distributed to two specific program cohorts, Systems and Biomedical, and an elective "society, technology, and values" course which contained multiple different engineering majors. We received a total of 53 participants from a variety of majors within engineering (see Table 1). The majority of participants were from three ma-

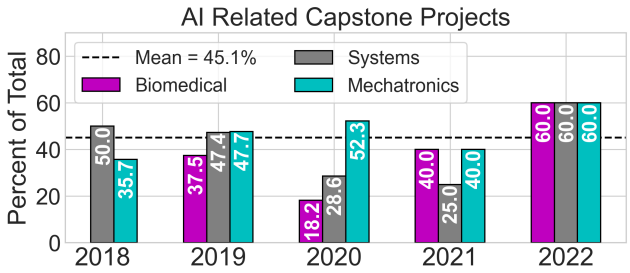

Figure 1: Percent of Systems, Biomedical, and Mechatronics Engineering capstone projects that are related to AI in each of the last five years.

jors: Systems (34.0%), Biomedical (30.2%), and Mechatronics (18.9%) Engineering. Other (17.0%) participants originated from a combination of Electrical, Software, and Chemical Engineering, to name a few. During the study, participants from Mechatronics (18.9%) and Other (17.0%) were enrolled in a "society, technology, and values" course. They received identical pre- and post-surveys at the beginning and end of the course (the same survey as Systems and Biomedical received once). However, there were limited students who completed both surveys (per unique codes), so we were unable to do paired data analysis. According to the course calendars for all engineering majors represented in this sample, all participants will have taken a mandatory first-year concepts and/or communication course that introduces them to engineering professionalization, which includes reviewing professional codes of engineering ethics. Participants in Systems will have taken a second-year human factors course and participants in Biomedical will have taken a third-year biomedical-specific ethics course.

Across the entire sample, 98% of participants had cooperative work terms, mostly in the software (35.1%), medical (11.5%), and robotics (8.1%) industries. In the past five years, 45.1% of capstone projects from Systems, Biomedical, and Mechatronics engineering cohorts contained AI related models, including 60% of projects in 2022 (Figure 1).[1] We see a 5.1% decrease in total AI related projects from 2020 to 2021. We suspect this may be a result of the COVID-19 pandemic, which made it difficult for students to meet and produce more labor intensive projects. By tracking the percentage of AI related capstone topics, we are able to see how 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.

Based on our participants' past curricular and work experiences, we can expect that they have had some exposure to these non-technical areas, in addition to a basic knowledge of artificial intelligence and machine learning.

Major	Responses	Year	% Total
Systems	18	3	34.0
Biomedical	16	3	30.2
Mechatronics	10	1 – 4	18.9
Other	9	1 – 4	17.0
Total	53	–	100.0

Table 1: Engineering major details for participants.

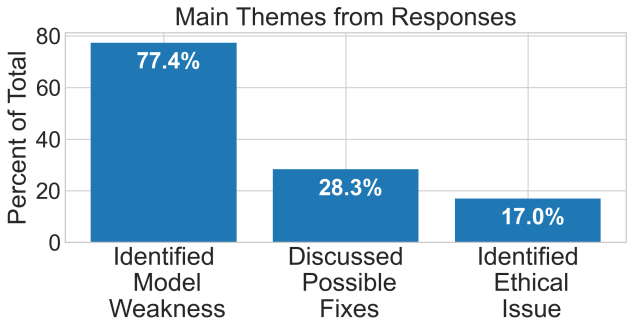

Figure 2: Percent of responses that identified the model's weakness, discussed possible methods to improve the model, and identified underlying ethical issues.

Results

Scenario Responses

In the short answer responses, we identified three themes: (1) participants identified a weakness or limitation in the facial recognition model; (2) participants discussed possible actions to improve the model; (3) participants identified how a biased model is a result of or could cause an ethical issue.

Identified Model Weakness The short answer question asks participants to "Please explain how you would respond in this scenario." In our analysis, we classified responses identifying the model weakness as instances where participants acknowledged that the imbalanced accuracy on the testing dataset was problematic. Our results showed that 77.4% of participants were able to identify this weakness (Figure 2 left bar), with the majority noting that the cause may be the distribution of the training dataset. This suggests that the participants were aware that classification accuracy imbalance is an issue and of how supervised learning models adopt the bias of the data they are trained on.

Discussed Possible Fixes In the same short answer space, only 36.6% of the number of participants that initially identified the model weakness (28.3% of all participants) discussed any method of how they could improve the model to mitigate bias (Figure 2 middle bar). While most participants identified a model weakness, the significant decrease in those that suggested model improvements suggests that the majority of the participant group chose not to respond by proposing a remedy. Of those participants that discussed

[1] https://uwaterloo.ca/capstone-design/project-abstracts

possible fixes, 73.3% blamed the issue on an insufficient dataset distribution. We cannot assume whether participants are familiar with other methods for balancing this distribution or modifying training techniques, as discussed in the Background, but in the Discussion section we elaborate on assumptions of dataset modification.

Identified Ethical Issues By identifying and expressing the importance of a potential ethical issue in this scenario, participants demonstrated their knowledge of and engagement with issues in facial recognition technology. Only 17.0% of participants expanded on how this problem could be resulting from or, if the model was deployed, could result in negative ethical implications. Some participants suggested the model could reinforce racism and discrimination:

- "The consideration of racial bias is extremely important as it's led to severe consequences for racialized communities in the past, and I think it's important to investigate racial disparities in the dataset used to build the model."
- "There needs to be individual reflection on how results like this in industry and academic research can lead to enforcing societal problems (racism and discrimination of many other forms)."

Other participants highlighted questions about details of the scenario, such as assuming the intended population:

- "Do I understand that this model is potentially racist? Is this project actually going to see implementation? Would I change it if it was implemented in the real world, as this could affect many individuals in society with its racism? Why is the intended population in the design not including darker skinned individuals? Why am I getting a high grade if this model is potentially racist?"
- "The least important consideration here is the fact that you have your intended population 'mostly' accounted for. Instead, the dataset should equally cover all demographics to ensure the fairest results are produced by the model."
- "Even though most of the intended population is accounted for, there are still other people that should be considered to increase inclusivity of the model."

Going beyond identifying the model weakness, these responses demonstrate more critical thinking about the purpose and scope of the AI model itself.

Ethics Curriculum and Co-op Responses

Participants were also asked to reflect on how ethics is included in other coursework and co-op contexts and any obstacles they have experienced in using ethical concepts.

One participant said there hasn't been any use for ethical reasoning in their technical courses, while another had it integrated into their capstone project:

- "My other courses are technical and there hasn't been a need to include any ethical reasoning."
- "I have applied concepts like these in the Machine Learning forth year capstone project I am working on such as looking into bias from the dataset we used and how we would be transparent with our future users and also how we would store their data to protect their privacy."

Participants were also asked if they had seen these concepts in their previous co-op placements:

- "I showed off [my knowledge of] these concepts and got a co-op placement."
- "I frequently work with data, privacy, and security concepts at my co-op workplace."

Though some participants have implemented these concepts, they also perceived obstacles to incorporating them into other courses, projects, or co-op placements:

- "There is often no opportunity to apply these concepts in school assignments."
- "Some obstacles could be the fact that projects at school are usually controlled environments where we aren't subjected to real world issues, such as algorithmic bias. It would be useful to have projects that do have these pitfalls in order to exercise our ability's to navigate them properly."
- "It hasn't happened before, but if my supervisor is largely dismissive of these ethical concerns then I would likely follow along to some extent."
- "A big obstacle is the fact that I'm a co-op student. Usually since I have this title I'm given less respect compared to other team members, so when I raise a concern usually they are weighted less compared to other (more senior) team members."

In summary, though many participants were able to identify and discuss how to mitigate the technical aspect of the issue in the scenario, few made connections with broader ethical or societal implications. Participants discussed the limited opportunities to apply ethical thinking in settings such as technical coursework and co-op placements, some obstacles being the lack of emphasis on ethics in their courses and their junior role in work placements.

Discussion

In their responses to the FRT scenario, most study participants linked the model weakness to insufficient dataset distribution and less than half discussed possible fixes, which included modifying or replacing the dataset to mitigate bias. Meanwhile, only 17.0% of participants noted how this issue could be resulting from or result in negative ethical implications such as discrimination. Nevertheless, the participants who noted potential ethical concerns raised critical questions about the purpose and implementation of the model. Teaching students to question the assumptions of a model's design and use should be a primary goal of any AI ethics course; the question is, how can educators best develop and instill this practice across the curriculum such that ethical inquiry is more embedded in the engineering workflow? A main obstacle observed by participants is the limited opportunities to practice thinking ethically in their technical courses and co-op workplaces. One participant suggested that "It would be useful to have projects that do have these pitfalls in order to exercise our abilit[ies] to navigate them properly."

Our results, in combination with evidence from our literature review on ethics education, suggest that sociotechnical pedagogical models would complement students' existing technical skills and enhance their ability to ask critical questions in mitigating complex ethical issues. Ideally, as noted by (Grosz et al. 2019; Tuovinen and Rohunen 2021; Fiesler, Garrett, and Beard 2020), a sociotechnical approach would be emphasized across the curriculum and not discussed in isolation from technical coursework.

To explore how our FRT scenario could be addressed with a sociotechnical framework, we consider approaches proposed by (Saltz et al. 2019) and (Krakowski et al. 2022). Saltz et al.'s question framework, designed for easy portability into technical ML modules, shows promise in supporting instructors who want to add ethics into their courses. It is practical for teaching alongside techniques for artificially modifying a dataset or how a model learns and asks questions that are applicable across different areas of AI. However, one limitation of this approach is its focus on ethical issues that can be addressed in the confines of their ML project and are not societal in nature – taken to mean that students are not explicitly prompted to more deeply consider the real world impact of their work.

Krakowski et al.'s framework, unlike Saltz et al.'s, is not connected with specific AI techniques, due to its initial application in high school curricula, and is thus less readily transferable to university courses. One advantage of Krakowski et al.'s framework is it contains broader questions about whether AI is an appropriate tool for the defined task, the data being used, the model's design, and how its output will impact the real world. This approach increases the potential for ethical inquiry, allowing for deeper engagement with different social contexts and concerns. This framework also resonates with responses to our study where participants questioned the intended population of the model and mentioned potential consequences for racialized communities.

As noted by (Raji et al. 2020a), striving for equal representation by adding more real samples of diverse individuals, for example, may contribute to greater privacy and consent issues and perpetuate discrimination of marginalized populations. Notably, of the participants who discussed possible fixes to the FRT scenario model, 73.3% blamed the issue on an insufficient dataset distribution and suggested to modify or replace the dataset. The implications discussed in (Raji et al. 2020a) and others is the type of ethical context that would be beneficial to discuss in a technical course because it makes clear the relevance of the social context to the technical solutions. Simply opting to modify or replace a dataset, though it might appear to address the issue on first glance, could elicit further undesirable consequences.

Connecting these two spheres of knowledge, as observed by (Tuovinen and Rohunen 2021), is crucial in signifying the importance of ethics to students. Shown by (Krakowski et al. 2022), the ability to formulate critical questions demonstrates a more formal understanding of ethical concerns.

This is not an exhaustive review of sociotechnical pedagogical approaches; nonetheless, we envisage that a combination of Krakowski et al. and Saltz et al.'s frameworks could produce very interesting and constructive AI ethics curricula that enhances students' abilities to address problems with complex sociotechnical considerations.

Limitations and Future Research

By surveying engineering undergraduates from across multiple subdisciplines, primarily Systems, Biomedical, and Mechatronics, we are not able to make generalizations about the kinds of AI education they have been exposed to. Our sample composition does not reflect all disciplines that utilize AI; however, their engagement with AI is demonstrated by 45.1% of capstone projects from 2018–2022 containing AI related models. Regardless of whether AI and machine learning is a prominent topic in their undergraduate curriculum, many students across engineering have been, and are likely to be, engaged with this technology. For this reason, we suggest that AI ethics be a required component in technical classrooms across different disciplines. One limitation is that a paired analysis for the elective course was not possible; as a result, we are unable to comment on whether the elective course had any influence in participant responses. To this end, we can only reference their experience in the mandatory first-year professionalism course, as with all other participants.

Our survey design contained open-ended prompts (e.g., "Please explain how you would respond in this scenario") that did not explicitly prompt participants to provide solutions or explain the ethical concerns of the scenario. As such, we cannot assume that any participants who did not elaborate on ethics are unaware of these concerns. Krakowski et al. surveyed their participants on a similar FRT scenario but instead used phrasing "What questions do you have?" in their design – appearing to cue students to apply the framework they learned (Krakowski et al. 2022); when measuring a specific learning outcome, future work may consider survey questions with similar framing. Placing our scenario in a curricular context was beneficial, as participants were familiar with the stakes and options when responding, but also a hindrance when participants dismissed the significance of addressing an issue that would not have "real world" impact because it is only coursework. Future work could explore using a co-op, workplace, or community-based context with various stakeholders.

As with any voluntary study, there may have been selection bias in who participated based on interest or experience.

Future research should continue to investigate findings on the various integrated ethics education methodologies; special attention should be paid to mitigating obstacles for and supporting instructors who want to embed more ethics in their technical teaching but lack resources to do so.

This study is part of a broader project dedicated to integrating ethics across the engineering and tech curriculum at our institution. These results will contribute to informing the study's next iteration focused on first-year students. From there, we aim to scaffold ethics outcomes throughout undergraduate engineering programs. Our next steps also include collaborating with more engineering instructors to implement, assess, and compare ethical frameworks and other pedagogical strategies, particularly those with a sociotechnical approach.

Acknowledgments

Research towards this paper was partially funded by the Waterloo Interdisciplinary Trailblazer Fund, National Sciences and Engineering Research Council of Canada (NSERC), Ontario Graduate Scholarship (OGS), and Waterloo Artificial Intelligence Institute (WAII) Scholarship. We thank members of the Trailblazer team, Heather Love, Marcel O'Gorman, Jennifer Boger, and Carter Neal for their ongoing support. We also thank Alexander Fleck for his feedback on this work.

References

Aguinis, H.; and Bradley, K. J. 2014. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational research methods*, 17(4): 351–371.

Antoniou, J. 2021. Dealing with emerging AI technologies: Teaching and learning ethics for AI. In *Quality of Experience and Learning in Information Systems*, 79–93. Springer.

Benjamin, R. 2019. Race after technology: Abolitionist tools for the new jim code. *Social forces*.

Brown, S.; Davidovic, J.; and Hasan, A. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, 8(1): 2053951720983865.

Cech, E. A. 2014. Culture of disengagement in engineering education? *Science, Technology, & Human Values*, 39(1): 42–72.

Chiusi, F. 2020. Life in the automated society: How automated decision-making systems became mainstream, and what to do about it. *Algorithm Watch*.

Cohen, L.; Precel, H.; Triedman, H.; and Fisler, K. 2021. A New Model for Weaving Responsible Computing Into Courses Across the CS Curriculum. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 858–864.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 113–123.

d'Entremont, A.; Shelling, W.; Pelletier, J.; and Gerrits, H. 2022. Developing and deploying an introductory equity curriculum for engineering. In *Canadian Engineering Education Association (CEEA-ACEG22) Conference*.

Erhan, D.; Courville, A.; Bengio, Y.; and Vincent, P. 2010. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 201–208. JMLR Workshop and Conference Proceedings.

Fiesler, C.; Garrett, N.; and Beard, N. 2020. What do we teach when we teach tech ethics? a syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–295.

Furey, H.; and Martin, F. 2019. AI education matters: a modular approach to AI ethics education. *AI Matters*, 4(4): 13–15.

Garrett, N.; Beard, N.; and Fiesler, C. 2020. More Than" If Time Allows" The Role of Ethics in AI Education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 272–278.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Green, B. 2021. The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing*, 2(3): 209–225.

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

Hamidi, F.; Scheuerman, M. K.; and Branham, S. M. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–13.

Hare, S. 2022. Technology Is Not Neutral: A Short Guide to Technology Ethics. *London Publishing Partnership Londres*.

Herkert, J. R. 2000. Engineering ethics education in the USA: Content, pedagogy and curriculum. *European Journal of Engineering Education*, 25(4): 303–313.

Hess, J. L.; and Fore, G. 2018. A systematic literature review of US engineering ethics interventions. *Science and engineering ethics*, 24(2): 551–583.

Hipp, C. 2007. An integrative approach to teaching engineering ethics. In *2007 Annual Conference & Exposition*, 12–223.

Hishiyama, R.; and Shao, T. 2022. Educational Effects of the Case Method in Teaching AI Ethics. In *World Conference on Information Systems and Technologies*, 226–236. Springer.

Hoffmann, A. L. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915.

Hoople, G. D.; and Choi-Fitzpatrick, A. 2017. Engineering empathy: a multidisciplinary approach combining engineering, peace studies, and drones. In *2017 ASEE Annual Conference & Exposition*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. PMLR.

Johnson, D. 1994. Who should teach computer ethics and computers & society? *Acm Sigcas Computers and Society*, 24(2): 6–13.

Kenfack, P. J.; Sabbagh, K.; Rivera, A. R.; and Khan, A. 2022. RepFair-GAN: Mitigating Representation Bias in GANs Using Gradient Clipping. *arXiv preprint arXiv:2207.10653*.

Krakowski, A.; Greenwald, E.; Hurt, T.; Nonnecke, B.; and Cannady, M. 2022. Authentic Integration of Ethics and AI Through Sociotechnical, Problem-Based Learning. In

Twelfth AAAI Symposium on Educational Advances in Artificial Intelligence.

McGinn, R. 2018. *The ethical engineer: Contemporary concepts and cases*. Princeton University Press.

Mittelstadt, B. 2016. Automation, algorithms, and politics—auditing for transparency in content personalization systems. *International Journal of Communication*, 10: 12.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11): 501–507.

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

Mozilla. 2007. The Mozilla Manifesto Addendum. The Pledge for a Healthy Internet. https://www.mozilla.org/en-CA/about/manifesto/. Accessed: 2022-09-06.

Mozur, Paul. 2019. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html. Accessed: 2022-09-06.

Mulvenna, M.; Boger, J.; and Bond, R. 2017. Ethical by design: A manifesto. In *Proceedings of the European Conference on Cognitive Ergonomics 2017*, 51–54.

Nasir, O.; Muntaha, S.; Javed, R. T.; and Qadir, J. 2021. Work in Progress: Pedagogy of Engineering Ethics: A Bibliometric and Curricular Analysis. In *2021 IEEE Global Engineering Education Conference (EDUCON)*, 1553–1557.

Noble, S. U. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

Perez, L.; and Wang, J. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020a. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.

Raji, I. D.; Scheuerman, M. K.; and 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*, 515–525.

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

Roncin, A. 2013. Thoughts on engineering ethics education in Canada. *Proceedings of the Canadian Engineering Education Association (CEEA)*.

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

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Taylor, L.; and Nitschke, G. 2018. Improving deep learning with generic data augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1542–1547. IEEE.

The Future of Life. 2017. Asilomar AI Principles. https://futureoflife.org/2017/08/11/ai-principles/. Accessed: 2022-09-06.

Truax, C.; Orchard, A.; and Love, H. A. 2021. The influence of curriculum and internship culture on developing ethical technologists: A case study of the University of Waterloo. In *2021 IEEE International Symposium on Technology and Society (ISTAS)*, 1–8. IEEE.

Tuovinen, L.; and Rohunen, A. 2021. Teaching AI Ethics to Engineering Students: Reflections on Syllabus Design and Teaching Methods. *Proceedings of the Conference on Technology Ethics*, 19–33.

Van Noorden, R. 2020. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834): 354–359.

Vecchione, B.; Levy, K.; and Barocas, S. 2021. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9.

Vraga, E. K.; Tully, M.; and Bode, L. 2020. Empowering users to respond to misinformation about Covid-19. *Media and communication (Lisboa)*, 8(2): 475–479.

Walczak, K.; Finelli, C.; Holsapple, M.; Sutkus, J.; Harding, T.; and Carpenter, D. 2010. Institutional Obstacles To Integrating Ethics Into The Curriculum And Strategies For Overcoming Them. In *2010 Annual Conference & Exposition*, 10.18260/1-2–16571. Louisville, Kentucky: ASEE Conferences.

Weber, J. 1992. Scenarios in business ethics research: Review, critical assessment, and recommendations. *Business Ethics Quarterly*, 2(2): 137–160.

Weiss, K.; Khoshgoftaar, T. M.; and Wang, D. 2016. A survey of transfer learning. *Journal of Big data*, 3(1): 1–40.

Zuboff, S.; Möllers, N.; Wood, D. M.; and Lyon, D. 2019. Surveillance Capitalism: An Interview with Shoshana Zuboff. *Surveillance & Society*, 17(1/2): 257–266.