Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025
Paper ID #47825

Expanding AI Ethics in Higher Education Technical Curricula: A Study on Perceptions and Learning Outcomes of College Students

Miss Indu Varshini Jayapal, University of Colorado Boulder
James KL Hammerman
Dr. Theodora Chaspari, University of Colorado Boulder

Theodora Chaspari is an Associate Professor in Computer Science and the Institute of Cognitive Science at University of Colorado Boulder. She has received a B.S. (2010) in Electrical & Computer Engineering from the National Technical University of Athens, Greece and M.S. (2012) and Ph.D. (2017) in Electrical Engineering from the University of Southern California (USC). Theodora's research interests lie in human-centered machine learning, affective computing, and biomedical health informatics. She is a recipient of the NSF CAREER Award (2021). She is serving as an Associate Editor of the Elsevier Computer Speech & Language and the IEEE Transactions on Affective Computing. Her work is supported by federal and private funding sources, including the NSF, NIH, NASA, IARPA, AFOSR, General Motors, and the Engineering Information Foundation.

Expanding AI Ethics in Higher Education Technical Curricula: A Study on Perceptions and Learning Outcomes of College Students

Indu Varshini Jayapal
induvarshini.jayapal@colorado.edu
University of Colorado Boulder
Boulder, CO, USA

James K. L. Hammerman
jim_hammerman@terc.edu
TERC
Cambridge, MA, USA

Theodora Chaspari
theodora.chaspari@colorado.edu
University of Colorado Boulder
Boulder, CO, USA

Abstract

The increasing integration of artificial intelligence (AI) into real-world applications requires researchers and developers to critically evaluate the ethical implications of their work and to work towards ethical design and implementation of AI technologies. However, AI ethics education receives limited attention in technical courses with many instructors feeling unprepared to teach ethics-related concepts. This gap risks fostering a workforce that develops AI technologies without adequately considering responsible and ethical practices, potentially leading to serious societal consequences. Here, we present results from a pilot curriculum that integrates various ethical topics related to AI into a graduate-level machine learning course. Activities include a combination of case studies, project-based learning, and critical classroom discussions on the ethical implications of AI systems design and deployment.

Two research questions guided the study: (RQ1) How do computer science graduate students perceive ethical issues in AI design and implementation before taking the class? (RQ2) How do these perceptions develop or change by the end of the AI course? This study employed a pre-post course survey method. Data were collected from 66 students enrolled in the Spring 2024 graduate machine learning course at the University of Colorado Boulder. The survey included 28 Likert scale questions addressing perception of ethics in diverse areas such as attention to ethical issues at various stages of design and implementation, fairness, accountability/ decision-making, transparency, and privacy. It also included open-ended essay questions about ethical dilemmas in high-risk AI applications, such as medical diagnosis and employee selection processes.

A combination of qualitative and quantitative analysis of the responses was used to identify the impact of our curriculum. We conducted factor analysis of the questionnaire items to understand (and simplify) their structure, finding four theoretically and empirically coherent factors, then used regression analyses to examine pre- to post- differences in these scores, and whether scores differed by demographic variables (gender, age, degree program, prior training in human subjects research). We also coded the students' responses to open-ended essay questions and identified themes derived from the responses, such as social / ethical / organization risks, human oversight, fairness, and privacy. To assess the complexity of students' ethical reasoning, we also evaluated the depth of discussion for each identified theme using a 4-point theme-depth scale (0-3).

The results indicate that while there is no significant pre- to post-assessment change in the factor scores derived from the 28-item questionnaire, students demonstrated a significant increase in the depth of their responses to the open-ended essay questions, with theme depth increasing from an average of 1.4 to 1.6. These are discussed alongside recommendations for future AI ethics curriculum design for computer science graduate students.

Keywords: Ethical AI, Ethical decision-making, Curriculum Development, Machine learning Curriculum, AI Fairness, Privacy, Explainability, Transparency.

1 Introduction

While artificial intelligence (AI) promises to improve our quality of life by automating tasks, advancing healthcare, and driving innovation, it also comes with risks that require careful ethical consideration [1]. The potential for AI to perpetuate or amplify biases in decision-making and infringe on privacy raise critical concerns about its responsible development. Additionally, the largely unchecked use of AI in consequential applications such as those pertaining to health, surveillance, and immigration decisions, can exacerbate social inequalities and infringe on fundamental rights [2]. These risks underscore the importance of embedding ethical principles, transparency, and accountability into AI development to ensure that its benefits are equitably distributed and do not come at the expense of societal well-being.

With mounting pressure for technology companies to consider societal impacts of AI in their systems and products, the demand of technical talent in AI-related roles continues to rise both in the public and private sectors. To meet this demand, universities have expanded their curricula to include multiple AI-focused courses [3, 4]. However, AI ethics training often receives little attention with many instructors feeling unequipped to teach ethics-related concepts with technical computing programs [5, 6]. As AI ethics becomes a key topic in public discourse, some universities have begun to add content on ethics and the responsible development of AI systems into their existing curricula [7].

The question on how to approach ethics in computing and engineering curricula is a long standing one. In early discussions of computing pedagogy, it was argued that ethics should be taught by social scientists and philosophers who possess the disciplinary expertise of the topic [8]. However, this usually results in standalone classes that tend to get marginalized and might sometimes ignore ethical issues that arise in the context of specific computing concepts and examples. To address these challenges, many educators and researchers advocate that ethics should be a necessary part of daily practice and should be integrated into technical computing and engineering courses in addition to the standalone ones [9]. In this way, students can engage with ethical issues as they arise naturally in the curriculum exploring key concepts in the responsible and ethical development and deployment of AI technologies, such as explainability, trustworthiness, and fairness [10]. Within the computing education research community, multiple approaches have been proposed for ethics pedagogy, including gamification [11], immersive theater [12], incorporation of science fiction [13], and use of codes of ethics [14].

We designed a pilot curriculum that integrates various ethical topics related to AI into a graduate-level machine learning course that attracts a large number of students. The curriculum was created in collaboration with the course instructor with expertise in machine learning and a STEM education expert with extensive experience in designing interdisciplinary curricula and educational interventions. The class activities included lectures where instructors discussed risks associated with AI and strategies for mitigating them, hands-on homework assignments to encourage reflection on the ethical implications of AI systems, and project-based work where students collaborated on human-centered AI problems with societal impact. To assess the curriculum's effectiveness, we conducted a pre/post evaluation measuring changes in students' perceptions of ethical issues in AI design and implementation. We used a mixed-methods approach, combining quantitative questionnaire data with qualitative insights gathered from open-ended essay questions to answer our two research questions: (RQ1) How do computer science graduate students perceive ethical issues in AI design and implementation before taking the class? (RQ2) How do these perceptions develop or change by the end of the AI course?.

2 Prior Work

Previous scholars have developed curricula focused specifically on ethics in AI [15] or have adopted an integrated approach, examining societal implications of algorithms alongside their technical structure and applications [16, 17, 18]. Courses that have introduced societal implications of AI to primary and middle school students have used lectures where instructors discuss risks associated with AI and corresponding mitigation methods [17] and project-based learning where students implement algorithms while reflecting on their societal implications [16]. Standalone courses on ethical and responsible AI have centered around concepts of safety, fairness, privacy, and ethical risks [19]. While these courses can have a significant impact, they are typically offered as electives and are not widely attended by the majority of undergraduate and graduate students.

At the higher education level, Saltz et al. [20] developed a framework of ten ethical questions for machine learning projects based on a systematic literature review. They also introduced course modules that integrate ethical questions into common ML assignments such as Linear Regression and Random Forest Classifier, finding that while ethics is part of the overall educational landscape in computing programs, it is rarely integrated into core technical ML courses. Recent research shows that ethics interventions in technical courses like Human Centered Computing [21] and Natural Language Processing [22] can successfully engage students without sacrificing core technical material. There has also been research focusing on designing online courses that teach ethics in AI. As part of this effort, a five-hour-long online AI ethics module combining case-based learning with reflective exercises was proposed [23]. These case studies addressed real-world ethical dilemmas, such as biased AI in autonomous vehicles, data misuse, and ethical challenges in AI-generated art. Students reflected on these cases to identify ethical issues and propose potential solutions.

Other initiatives have designed activities prompting students to identify stakeholders of AI systems and evaluate how these stakeholders might be positively or negatively affected [11, 24]. These reflections were enhanced through gamification - offering an interactive platform for peer discussions and scenario exploration. Researchers have also incorporated science fiction stories and short articles about AI systems, encouraging students to apply major ethical theories not only as evaluative tools but also as frameworks for identifying problems and exploring solutions from diverse perspectives [5], as well as an art-based approach in which students build and manipulate AI systems to create digital media [25]. Additionally, a noteworthy research effort brought together computer science and theater students to co-create an immersive theater production, blending sci-fi drama with interactive art installations to build an explorable world where audiences reflect on the future of data in society [12].

Recent work has explored co-designing AI ethical frameworks with middle and high school students to make these frameworks more relevant to their experiences with technology [26]. Findings from these studies reveal that existing AI ethics frameworks, such as the White House Blueprint for AI [27], are often too broad and fail to resonate with students, underscoring the need for tailored, relatable approaches. At the higher education level, a realist review by Padiyath [28] identified several factors that affect student acceptance of ethical interventions, including preconceived notions about ethics in computing and prioritization of technical concepts over ethics, suggesting that successful integration requires careful attention to course design and classroom dynamics.

The contributions of our research, compared to prior work in addressing some of the gaps, are two-fold: (1) While several valuable approaches for AI ethics education exist for both K-12 and higher education settings [20, 21, 22], our work stands apart in its multifaceted integration of AI ethics concepts into core technical content. Unlike previous efforts that primarily added reflection questions while building AI [22] or incorporated guest lectures and reflection questions into projects [21], we integrate ethical considerations directly into homework assignments, encouraging students to build more ethical AI systems than what typical technical ML assignments require. This includes introducing students to alternative evaluation metrics such as "equality of opportunity" and guiding them to explore and implement these metrics in their coursework; and (2) Many existing initiatives for college students [19, 25] primarily focus on standalone AI ethics courses or brief modules spanning just a few hours [23], which may not reach the majority of technical AI students. In contrast, our study integrates AI ethics directly into an existing technical AI course that attracts a large number of computer science graduate students, providing them with ethics education that engages with the complex technical and ethical challenges they will face professionally. We believe this integrated approach increases the likelihood that future AI practitioners will receive substantive ethics training that will influence their professional practice.

3 Methods

3.1 Curriculum Design

The curriculum was designed by the course instructor, who has seven years of experience teaching the machine learning course, in consultation with a STEM education expert with extensive expertise in creating interdisciplinary curricula and educational interventions. Core aspects of AI trustworthiness and ethics, such as explainability, interpretability, fairness, privacy [10], were woven into various parts of the course including lectures, homework assignments, and a mini-project. Please refer Table 1 for more details.

Technical Concepts	Assignments	Ethical Concepts
Introduction to Machine Learning K-Nearest Neighbor	Homework 1: Predict breast cancer survival based on tabular data of tumor characteristics and patient demographics	• Discuss implications of algorithmic deployment and generalizability in real-life healthcare settings. • Explore algorithmic bias with respect to socio-demographic groups.
Data Pre-processing (Non-)Linear Regression Logistic Regression	Homework 2: Estimate water salinity based on biochemical oceanographic measurements	• Discuss implications of algorithmic deployment to increase our understanding of climate change. • Reflect on the interpretability of the resulting AI model and how it can be used for decision-making.
Neural Networks	Homework 3: Image-based object recognition	• Lecture on AI explainability methods and AI trustworthiness. • Apply explainability methods in the context of object recognition. • Conduct and discuss demo on human-AI interaction for decision-making.
Decision Trees and Random Forests	Homework 4: Predict one's hireability based on acoustic and physiological measures captured in a job interview	• Discuss implications of deploying AI systems for employee selection.
Feature Selection and Transformation Clustering Ensemble Learning	Team project: Detect depression level based on speech	• Lecture on algorithmic bias and fairness. • Assess AI privacy leaking risks and socio-demographic bias. • Experiment with bias mitigation algorithms. • Discuss implications of AI systems in fair and ethical treatment.

Table 1: Curriculum of machine learning course interweaving technical and ethical concepts.

Lectures included dedicated discussions on the societal implications of AI, exploring topics such as explainability, trustworthiness, and fairness, which are key concepts in the responsible and ethical development and deployment of AI technologies [10]. These lectures were carefully aligned with the technical content of the course to ensure relevance and contextual understanding. For instance, the lecture on AI explainability was presented in the context of addressing challenges associated with "black-box" models like neural networks.

Four homework assignments were developed to integrate core technical material from the course while incorporating questions on trustworthy AI and the responsible design and deployment of AI technologies. These assignments tackled problems with significant societal impact, such as those in health, environmental science, and education/training, offering a rich foundation for exploring the ethical dimensions of AI models. Additionally, a mini-project conducted in student teams was assigned at the end of the semester. This project provided students with hands-on experience working with a real-world, human-centered dataset, encouraging them to reflect on the real-life consequences of AI system design. It also tasked them with devising mitigation strategies to address the leaking of personally identifiable information and algorithmic bias.

3.2 Participant Recruitment

The participants of the study were enrolled in two sections of a machine learning course (CSCI 5622 - 001/002) at the University of Colorado Boulder between January 16, 2024, and May 7, 2024. The 001 section was conducted in person, while the 002 section was held online. All students followed the curriculum outlined earlier in this paper. To evaluate the impact of the curriculum, students were invited by the course evaluator to complete two surveys: one administered in the first week of the class and another in the last week of the class to measure pre/post differences. Each survey took approximately 30 minutes to complete and was administered via Qualtrics. Participation in the surveys was incentivized with bonus points for the course. Students who opted not to participate in the research could either complete an alternative activity to earn the same bonus points or choose not to engage in any additional activity.

Out of the 86 students the survey was sent to, 38 responded to the pre survey and 66 students responded to the post survey, with 33 participants who responded to both the pre and post survey for the class. The group with both pre and post data were 64% male, 33% female, and 3% who identified as non-binary. They identified their race and ethnicity as 36% White, 64% Asian, 3% Hispanic or Latino, and 3% who preferred not to self-identify. Average age was 24.9 (SD =3.1) with a range from 20 to 32 years old. The class included 61% of students who are English-speakers, 48% who speak Hindi, Tamil, Telugu or other south Asian languages, and 3% who speak Korean. Almost all (91%) were graduate students, 15% were first generation college students, and 21% said they had received CITI training in working with human subjects in research. The students were distributed across disciplines as follows: 73% in Engineering and Applied Sciences, 24% in the Graduate School, 18% in Arts and Sciences, and 3% in multidisciplinary graduate programs.

3.3 Measures

The survey was designed to seek the students' perceptions and views on ethical considerations in the design of AI and its applications to society. Each survey consisted of two main components. First, as shown in Table 2, there were 28 statements about ethical design and implementation of AI, which students rated using a Likert scale from -3 (Strongly Disagree) to +3 (Strongly Agree). Second, as detailed in Table 3, there were 3 open-ended essay questions that presented situations involving AI technologies and asked students to identify potential concerns with those implementations. Demographic and background information (age, gender, race and ethnicity, primary language, degree level and college, prior CITI training in addressing issues of human subjects in research) were also collected. The pre and post surveys were the same except the post-class survey also included an additional component: a helpfulness score, where students were asked to rate each ethics-related activity. These helpfulness scores were collected using a 4-point scale (0 = Neutral to +3 = Strongly helped) for each curriculum component, including the lectures on AI explainability and fairness, the team-based mini-project, and the ethics-related homework questions.

3.4 Factor Analysis of Questionnaire Items

The 28 items of the questionnaire were designed to reflect important elements of ethical AI design and implementation – privacy and use of personal information, human control over AI decisions, interpretability and transparency of results, equitable benefits from AI, and overall attention to addressing harms and risks throughout the design and implementation of AI technologies. We conducted an exploratory factor analysis to reduce the dimensionality of the response data and to see whether items intended to load on similar underlying factors correlated with one another. A very simple structure (VSS) analysis [29] found four factors, as described in Table 2 with the corresponding questionnaire items. There was one item that did not load on any of these four factors, and two other items that loaded equally (and not very highly) on three of the four factors. These items are listed at the bottom of Table 2 and were deleted from further analysis.

VSS Factors	Questionnaire Item	Weight
Understand Control - Humans should understand the AI output and control its use	It's OK if people who use or are impacted by AI technologies do not understand how they make decisions.	Negative
	Users don't need to understand how AI technologies work, as long as they trust their results.	Negative
	How AI technologies make decisions should be clear to all those who use them or are impacted by them.	
	It's OK if only experts understand the output from AI technologies.	Negative
	Humans should always review the decisions the AI is making.	
	As long as an AI technology has been thoroughly evaluated, it is OK if it provides an output that does not make sense to its users.	Negative
	Humans should have the last word over the decisions that the AI is making.	
Trust in Technology - A mix of statements giving unquestioned trust in AI technologies and its suggestions, disregarding privacy concerns, and accepting benefits for some but not all social groups	Once an AI technology is launched, nothing should be done about it if there's evidence that it is being used to violate people's privacy.	
	There are times when it's OK for an AI technology to disclose people's personal information without permission.	
	An AI technology is ethical as long as it benefits some social or economic groups, even if it does not benefit everyone.	
	An AI technology that mostly benefits privileged social and economic groups is still ethical.	
	The most important thing to think about in designing AI technology is that it works.	
	Human control over AI should occur only at the design stage.	
	People should always follow the suggestions/decisions AI makes.	
	Results from AI shouldn't be overwritten by humans.	
	An AI technology should never use people's personal information without permission.	Negative
	An AI technology should be designed not to disclose people's personal information.	Negative
Ethical Design - Statements expressing a concern for the ethical impacts of AI technology throughout its lifespan	An AI designer's responsibility includes addressing both intentional and unintentional harm to humans.	
	It is important to think about ethical issues while designing AI technology.	
	It is important to review potential ethical impacts of an AI technology before launching it.	
	An AI designer should address how deployment of the AI might harm the environment, infrastructure, and/or non-human living beings.	
	AI designers must always work to address the risks arising from deployment of the AI in society.	
	It is important to review the actual ethical impacts of an AI technology after it's been in use for some time.	
AI SES Impacts - Concern for equal benefits to all socio-economic (SES) groups	AI technologies should benefit all social and economic groups equally.	
	AI technologies should have roughly the same impact on all social and economic groups.	
Deleted Items	Human users are ultimately responsible for the impact of decisions even if based on suggestions from AI technologies.	
	Users only need to understand how an AI technology accomplishes the purpose they have in mind.	
	Users should always be helped to understand all that AI technologies can do.	

Table 2: Factors resulting from the very simple structure (VSS) exploratory factor analysis with the corresponding items of the questionnaire.

Scores for each of these factors were determined for each participant at both pre- and post-assessments by averaging the ratings for all included items. This preserved the original Likert scale, so the resulting factor scores are on the –3 = Strongly disagree to +3 = Strongly agree scale. Paired and unpaired t-tests were performed to identify significant differences in factor scores between the pre- and post-assessments, both within a person and across the overall group, respectively. A structured regression analysis was conducted to examine the impact of participants' demographic characteristics on their factor scores, using several versions of the following equation:

$$\text{Factor Score} \sim \text{Survey Time (pre/post)} + \text{Demographic Variable} \tag{1}$$

3.5 Analysis of responses to open-ended essay questions

The three open-ended essay questions asked participants to discuss their questions and concerns about the design and implementation of AI systems across different domains in healthcare and employment selection (Table 3).

ID	Question
COVID	A machine learning (ML) algorithm has been designed to assist radiologists with estimating the level of damage COVID-19 has caused to patients' lungs. This can help the physician in prescribing an appropriate medication and treatment plan for the patients. The ML algorithm has been trained on a large set of X-ray images that were captured from patients in a private hospital in [it City removed for double-blinded submission]]. You are reviewing the algorithm that was trained with this data. The algorithm will be used by radiologists in a variety of settings with a range of experience and expertise. What questions / concerns / comments would you have in regards to the way the ML system has been designed and would be used?
Mental Health	Prior studies have found that certain kinds of degradation in mental health, such as depression, can be detected from individuals' vocal patterns, such as intonation (pitch), loudness, and prosody (rhythm or emphasis). A team of scientists is working on a system that will record users' voices through smartphones during people's daily lives. The recorded data will be sent to a server where a machine learning (ML) algorithm will be used to detect mental health degradation. The decision of the algorithm will be sent to the users' primary physician who will provide further guidance. You are reviewing the algorithm. What questions / concerns / comments would you have in regards to the way the ML system has been designed and would be used?
Employment	A big tech company has designed a machine learning (ML) system that will automatically sort candidates for a software engineer position based on their CV. The algorithm was trained on a set of CVs from demographically and academically diverse candidates. The tech company only plans to invite the top 10 candidates identified by the algorithm to interview for an open position. You are reviewing the algorithm. What questions / concerns / comments would you have in regards to the way the algorithm has been designed and would be used?

Table 3: Open-ended essay questions administered in pre and post surveys.

To code the responses, we used grounded theory methods to develop a thematic coding framework with seven categories addressing ethical issues arising in student responses. The coding themes were informed by existing principles of trustworthy AI [10]. One coder reviewed the responses to the questions multiple times to become familiar with the content of the responses. Following that, the coder identified significant phrases and sections of text and assigned descriptive codes to them. After an initial pass, the coder looked for patterns and similarities across the codes and grouped similar ideas together into overarching concepts. This resulted in a total of seven themes, as defined in Table 4.

Abbreviation	Theme Title	Theme Description
PosSocImpact	Positive Societal Impact	Discussions about how AI can address societal challenges and contribute positively to society.
SocOrgEthRisks	Societal/ Ethical/ Organizational Risks	Discussions on the risks associated with widespread AI implementation, including ethical concerns, regulatory compliance, and potential adverse effects on individuals or organizations.
Fairness	Fairness	Concerns about the representativeness of training data and potential socio-demographic biases in AI models.
PrivSecurity	Privacy and Security	Concerns on user privacy, consent for data use, risks arising from adversarial attacks, and misuse of personal information.
InterpExplnTransp	Interpretability/ Explainability/ Transparency	Discussions on the importance of making AI algorithms interpretable, explainable, and transparent to both technical and non-technical audiences.
Human Oversight	Human Oversight	Concerns regarding the reliance on AI for autonomous decision-making and the necessity of human involvement to ensure accountability and safety.
AlgDevtTech	Algorithmic Development and Technical Challenges	Discussions about the technical challenges related to designing, implementing, and testing AI systems.

Table 4: Description of the themes that were used for coding participants' responses to open-ended essay questions.

Participants' responses were categorized into one or more themes, as responses often addressed multiple dimensions, such as ethical risks and privacy concerns. To capture learning outcomes on ethical and responsible AI more effectively, the complexity and rigor of participants' discussions on each theme–referred to as theme depth–were also assessed. Theme depth was measured using a 4-point scale adapted from Baker-Brown et al.'s conceptual/integrative complexity framework [30]. This scale assesses the sophistication of students' engagement with ethical considerations in AI:

- **No mention (0):** The ethical theme is completely absent from the response.
- **Superficial mention (1):** The ethical issue is briefly acknowledged without substantive discussion. For example, a student might simply state "privacy is a concern" without explaining why or how it applies to the AI system in question.
- **Detailed description (2):** The ethical issue is clearly described with supporting context or examples. At this level, students identify specific ethical dimensions of the AI system and provide reasoned explanations of potential concerns.
- **Complex analysis (3):** The student thoroughly analyzes the ethical implications, discussing potential tensions, tradeoffs, or interdependencies between different ethical considerations. For instance, a student might examine how improving an AI system's accuracy might come at the cost of transparency or privacy.

This adapted framework allowed us to systematically evaluate the complexity of students' ethical reasoning while focusing on the depth of their critical thinking about AI ethics rather than simply the breadth of themes mentioned. The aforementioned depth scale was applied to each identified theme when coding the responses to the three open-ended questions. The coding process began with the first author (Coder 1) independently coding all student responses to the open-ended essay questions. To assess the reliability of the coding scheme, the second (Coder 2) and third (Coder 3) authors subsequently coded a randomly selected subset comprising 15% of the responses (i.e., 48 responses total per coder pair). The inter-coder agreement percentages for theme flagging were generally high (Table 5), with averages of 89.29% between Coder 1 and 2, and 87.80% between

Coder 1 and 3. Agreement on depth ratings was moderately high, averaging 76.79% between Coder 1 and 2, and 75.89% between Coder 1 and 3. Among the themes, Fairness showed the lowest agreement in depth coding, with percentages of 52.08% between Coder 1 and 2, and 50.00% between Coder 1 and 3. Algorithmic development depicted the second lowest agreement, with percentages of 58.33% between Coder 1 and 2, and 54.17% between Coder 1 and 3. We conducted a linear regression analysis, controlling for theme, to see whether ratings differed significantly by coder. Results indicated significant differences between coders when all themes were included; however, these differences were no longer significant when the Fairness theme was excluded from the regression model. Therefore, the Fairness theme was omitted from subsequent analysis.

Theme	**Theme presence**		**Theme depth**	
	Coder 1 - Coder 2	**Coder 1 - Coder 3**	**Coder 1 - Coder 2**	**Coder 1 - Coder 3**
Positive Societal Impact	100.00%	95.83%	97.92%	95.83%
Societal/Ethical/Organizational Risks	79.17%	85.42%	75.00%	81.25%
Algorithmic Development/Technical Challenges	79.17%	77.08%	58.33%	54.17%
Fairness	87.50%	79.17%	52.08%	50.00%
Privacy/Security	97.92%	100.00%	85.42%	85.42%
Interpretability/Explainability/Transparency	91.67%	89.58%	89.58%	79.17%
Human Oversight	89.58%	87.50%	79.17%	85.42%
Average	**89.29%**	**87.80%**	**76.79%**	**75.89%**

Table 5: Inter-coder agreement percentages for annotating the presence and depth of each theme.

The average number of themes and theme depth across all questions was computed for each participant for the pre- and post-assessment. Paired and unpaired t-tests were performed to identify significant differences in the number of themes and theme depth between the pre- and post-assessments, both within a person and across the overall group, respectively. We further computed Pearson's correlation coefficient between the theme depth and factor scores assessing potential associations between the concepts measured by the questionnaire and those captured in the open-ended questions. To check for patterns in whether a specific essay or a specific theme predicted overall significant changes in depth between pre and post-assessment, we restructured the data so each non-zero depth rating was associated with the participant, survey time (pre/post), essay, and theme type, which resulted in a total of 720 samples (people-time-essay-themes). We used a structured regression with the Tukey Honest Significant Differences procedure to account for multiple comparisons, as follows:

$$\text{Theme Depth} \sim (\text{Survey Time (pre/post)} * \text{Essay}) + (\text{Theme} * \text{Essay}) \tag{2}$$

4 Results

4.1 Helpfulness of ethics curriculum components

The average helpfulness score provided by students across all class activities was 2.41 (out of 3). The lecture dedicated to AI explainability received the highest score ($M = 2.47$, $SD = 0.79$), followed by the team-based mini-project ($M = 2.44$, $SD = 0.79$) and the lecture on fairness ($M = 2.42$, $SD = 0.77$). The ethics-related questions integrated in the homework assignments were rated the least useful ($M = 2.30$, $SD = 0.93$). No significant differences among the types of activities in terms of helpfulness were observed. The overall distribution of student responses are also shown in Figure 1.

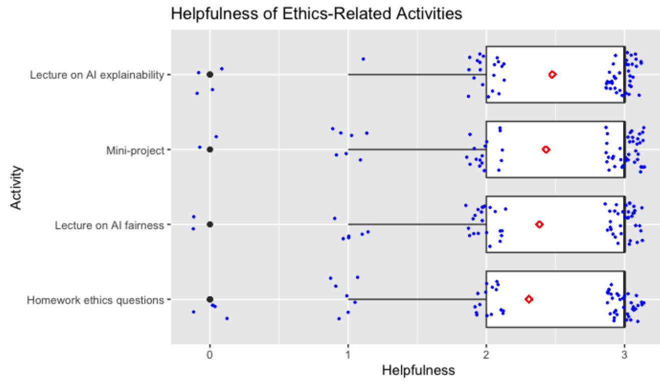

Figure 1: Distribution of usability scores of the different curriculum ethics components

4.2 Factor scores derived from Likert scale ratings

The distribution of factor scores for pre- and post-assessments is summarized in Table 6.

Factor	Pre ($N = 38$)		Post ($N = 66$)	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Understand Control	1.055 (1.043)	1.14 (.64, 1.57)	1.426 (1.043)	1.43 (.75, 2.29)
Tech First	-1.571 (0.661)	-1.70 (-2.08, -1.20)	-1.446 (0.980)	-1.75 (-2.13, -1.00)
Ethical Design	2.494 (0.536)	2.67 (2.17, 2.96)	2.382 (0.663)	2.50 (2.00, 3.00)
SES Impacts	0.947 (1.515)	1.25 (0.00, 2.00)	1.326 (1.445)	1.50 (0.13, 2.50)

Table 6: The distribution of factor scores across participants for the pre- and post-assessment.

There were no significant differences pre to post on any of the four factors whether looking at within person change or overall group differences. Most demographic and background characteristics (gender, degree program, prior training in work with human subjects) also did not predict these scores. There were significant age related differences for the Understand Control factor, with older students agreeing more than younger students ($\beta = 0.065$, $t = 2.22$, $df = 98$, $p = .029$), and those whose primary language is not English scoring higher than English speakers ($\beta = 0.45$, $t = 2.15$, $df = 101$, $p = .034$), though non-English speakers tend to be older so these two findings are confounded. We found that the factor scores tended to cluster together, but did not correlate with the essay depth ratings (Table 7), even when we excluded the algorithmic development theme (i.e., AlgDevtTech) from computing the average theme depth, so we decided to analyze them separately.

r (p)	**Understand Control**	**Tech First**	**Ethical Design**	**SES Impacts**
Tech First	-0.251 (.010)			
Ethical Design	0.194 (.048)	-0.302 (.002)		
SES Impacts	0.237 (.015)	-0.111 (.263)	0.149 (.132)	
Average Theme Depth	0.122 (.219)	0.039 (.698)	0.054 (.587)	-0.122 (.219)
Average Theme Depth excluding AlgDevtTech theme	0.139 (.159)	0.017 (.863)	0.159 (.108)	-0.047 (.635)

Table 7: Pearson's correlation coefficients among the different factor scores, as well as the factor scores and the theme depth.

4.3 Number of themes and theme depth of open-ended questions

Participants discussed on average 1.69 ($SD = 0.83$) of the considered themes in the beginning of the class and 1.74 ($SD = 1.00$) themes at the end of the class. Figure 2 shows the boxplots of the distribution of theme depth ratings for the pre- and post-assessment (left), as well as the individual difference scores (right) – with individual points jittered vertically but not horizontally. There are significant differences pre- to post- in the theme depth ($M_{\text{pre}} = 1.40$, $M_{\text{post}} = 1.60$, $M_{\text{diff}} = 0.21$, $t = 2.09$, $df = 96$, $p = .040$), though not in the average number of themes mentioned. Within person difference depth scores (subtracting pre- from post-) are also statistically significant (M_{diff} = 0.285, $t = 2.55$, $df = 32$, $p = .016$).

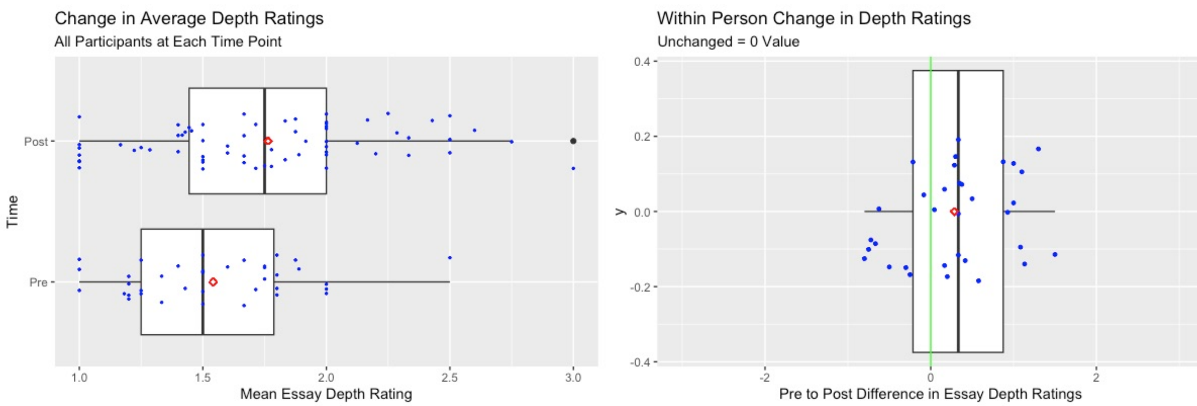

Figure 2: Boxplots showing changes in theme depth ratings. The left panel displays the overall distribution of theme depth ratings at pre- and post-assessment, where each point represents a student's average depth across all essays. The right panel shows individual student growth, with positive values indicating improvement in theme depth from pre- to post-assessment. Red diamonds indicate mean values in both panels. The green vertical line at zero in the right panel serves as a reference point separating positive from negative changes.

Depth ratings differ pre- to post- (M_{diff} = .22, $p = .012$). Depth ratings on the AlgDevtTech theme are higher than other themes (range of .27 to .70 points higher). If we exclude the AlgDevtTech theme because it mostly identifies a concern for technical issues related to creating the described AI technologies (resulting $N = 354$), there are still significant pre- to post-differences in the depth of ratings ($\beta = 0.20$, $t = 3.09$, $df = 350$, $p_{\text{adj}} = .006$) with ratings for the COVID essay lower than on the Employment essay ($\beta = 0.19$, ns, $p_{\text{adj}} = .08$) and on the Mental Health essay ($\beta = 0.28$, $t = 3.29$, $p_{\text{adj}} = .003$).

5 Discussion

Computer science graduate students at the beginning of a class on AI design have generally supportive views about ethical components (RQ1). They strongly agree that ethical issues should be addressed in design and implementation ($M = 2.5$ on a 3 point scale), they slightly agree that humans should understand AI output and control its use ($M = 1.1$) and that AI should benefit all socioeconomic groups equally ($M = 0.9$), and they disagree somewhat that AI designers should unquestioningly trust AI and ignore privacy issues if that helps AI work ($M = -1.6$). In more open-ended essays describing questions and concerns about ethical situations, they mention a little more than 1 non-technical theme per essay ($M = 1.13$), and either mention these superficially or describe them with a bit of detail ($M_{\text{depth}} = 1.53$). Prior to the class, the depth of students' focus on algorithmic and technical topics was consistently higher than the depth of their focus on ethical issues. Given that the majority of students majored in Computer Science, this tendency may be attributed to prior educational experiences, which likely emphasized technical problem-solving over critical reflection on ethics and societal impacts.

By the end of the AI course (RQ2) results indicate no significant difference in students' ethical views as measured by the rating scales, although there is a marginally significant increase in their sense that humans should understand and control the use of AI. There is also a significant improvement in students' understanding and articulation of ethical issues in more open-ended contexts. By the end of the course, participants discussed a comparable number of ethically related themes ($M_{\text{diff}} = 0.11$, $t = 0.67$, $df = 102$, $p = .51$) but exhibited significantly greater depth in their responses ($M_{\text{diff}} = 0.21$, $t = 2.39$, $df = 101$, $p = .019$) compared to the beginning of the class. Overall, the increased depth in essay responses may highlight a development in students' ability to consider ethical issues in AI systems, moving beyond technical concerns to address societal implications. However, the factor scores derived from the 28-item questionnaire did not show a significant improvement pre- and post-assessment and only Understand Control and SES Impacts moved in the hoped for direction, though there may have been a ceiling effect for Ethical Design since agreement with these statements started off quite high. This lack of statistically significant change in factor scores may reflect the specific focus of the course content and nature of assignments, indicating that the methods in the curriculum may have succeeded in deepening students' reflective understanding of specific ethical issues through practical applications, but it may not have been as effective in altering their overall beliefs about ethical AI.

Despite the promising results, this study has the following limitations. First, since the participants were primarily computer science graduate students, the findings cannot be generalized to broader populations such as graduate students from other disciplines or engineering students. Second, the relatively small sample size limits our ability to explore differences in how the course impacts different groups of students based on their demographic characteristics and prior experiences. Additionally, we were unable to determine whether any students used generative AI tools to complete their essay responses. As part of our future work, we aim to incorporate automated tools and collect student self-reports to address this issue. The reliance on student self-reports about their beliefs and open-ended essays might not fully capture the breadth of students' understanding. Adding assessments such as case study analyses, design projects, or peer evaluations could offer a more comprehensive view of students' grasp of ethical AI issues. Lastly, the curriculum could be enhanced through future revisions, including role-playing or debates on controversial AI ethics topics, expanding group projects to foster collaborative learning, and integrating more real-world case studies from fields like healthcare and criminal justice to demonstrate the tangible societal impacts of AI systems.

6 Conclusion

In conclusion, this study highlights the potential of integrating ethics-focused curricula into technical AI courses to enhance students' awareness and articulation of ethical considerations in AI design and implementation. While quantitative measures showed no significant change in factor scores, the qualitative analysis revealed notable improvements in students' ability to critically engage with ethical dilemmas. These findings underscore the importance of using diverse teaching methods, such as dedicated case studies and project-based hands-on activities, to foster deeper understanding and engagement with ethical issues. Future iterations of this curriculum can build on these results by incorporating more comprehensive assessments and refining activities to better address gaps in students' ethical reasoning skills.

7 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2046118. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This research has been approved by the Institutional Review Board (IRB) of the University of Colorado Boulder (IRB Protocol Number: 23-0735).

References

[1] B. Cheatham, K. Javanmardian, and H. Samandari, "Confronting the risks of artificial intelligence," *McKinsey Quarterly*, vol. 2, no. 38, pp. 1–9, 2019.

[2] M. Kalvaitytė, "Unregulated Negative Impacts of AI: Mixed Methods Analysis of Feedback Responses to the EU AI Act Proposal," Master's thesis, Sciences Po, France, 2022.

[3] J. Southworth, K. Migliaccio, J. Glover, D. Reed, C. McCarty, J. Brendemuhl, A. Thomas, *et al.*, "Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in AI literacy," *Computers and Education: Artificial Intelligence*, vol. 4, p. 100127, 2023.

[4] H. S. Kim, D. Kim, S. I. Kim, and W. J. Lee, "Analysis of the Current Status of the AI Major Curriculum at Universities Based on Standard of AI Curriculum," *Journal of The Korea Society of Computer and Information*, vol. 27, no. 3, pp. 25–31, 2022.

[5] E. Burton, K. Clayville, S. A. Doore, M. S. Kirkpatrick, and M. Goldweber, "Managing Authority When Teaching Computing Ethics," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, pp. 1523–1524, 2024.

[6] N. Garrett, N. Beard, and C. Fiesler, "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, 2020.

[7] C. Fiesler, N. Garrett, and N. Beard, "What do we teach when we teach tech ethics? A syllabi analysis," in *Proceedings of the 51st ACM technical symposium on computer science education*, pp. 289–295, 2020.

[8] D. Johnson, "Who should teach computer ethics and computers & society?," *Acm Sigcas Computers and Society*, vol. 24, no. 2, pp. 6–13, 1994.

[9] B. J. Grosz, D. G. Grant, K. Vredenburgh, J. Behrends, L. Hu, A. Simmons, and J. Waldo, "Embedded EthiCS: integrating ethics across CS education," *Communications of the ACM*, vol. 62, no. 8, pp. 54–61, 2019.

[10] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy AI: From Principles to Practices," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–46, 2023.

[11] S. Ali, V. Kumar, and C. Breazeal, "AI audit: a card game to reflect on everyday AI systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 15981–15989, 2023.

[12] M. Skirpan, J. Cameron, and T. Yeh, "Quantified self: An interdisciplinary immersive theater project supporting a collaborative learning environment for cs ethics," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pp. 946–951, 2018.

[13] E. Burton, J. Goldsmith, and N. Mattei, "How to teach computer ethics through science fiction," *Communications of the ACM*, vol. 61, no. 8, pp. 54–64, 2018.

[14] M. J. Wolf, D. Gotterbarn, and M. S. Kirkpatrick, "ACM code of ethics: Looking back and forging ahead," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, pp. 801–802, 2019.

[15] S. Ali, B. H. Payne, R. Williams, H. W. Park, and C. Breazeal, "Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education," in *International workshop on education in artificial intelligence k-12 (eduai'19)*, vol. 2, pp. 1–4, mit media lab Palo Alto, California, 2019.

[16] R. Williams, S. Ali, N. Devasia, D. DiPaola, J. Hong, S. P. Kaputsos, B. Jordan, and C. Breazeal, "AI+ ethics curricula for middle school youth: Lessons learned from three project-based curricula," *International Journal of Artificial Intelligence in Education*, vol. 33, no. 2, pp. 325–383, 2023.

[17] H. Zhang, I. Lee, S. Ali, D. DiPaola, Y. Cheng, and C. Breazeal, "Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study," *International Journal of Artificial Intelligence in Education*, vol. 33, no. 2, pp. 290–324, 2023.

[18] T. K. Chiu, H. Meng, C.-S. Chai, I. King, S. Wong, and Y. Yam, "Creation and evaluation of a pretertiary artificial intelligence (AI) curriculum," *IEEE Transactions on Education*, vol. 65, no. 1, pp. 30–39, 2021.

[19] A. Alam, "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 *Intelligent Communication Technologies and Virtual Mobile Networks*, pp. 879–894, Springer, 2023.

[20] J. Saltz, M. Skirpan, C. Fiesler, M. Gorelick, T. Yeh, R. Heckman, N. Dewar, and N. Beard, "Integrating ethics within machine learning courses," *ACM Trans. Comput. Educ.*, vol. 19, Aug. 2019.

[21] M. Skirpan, N. Beard, S. Bhaduri, C. Fiesler, and T. Yeh, "Ethics education in context: A case study of novel ethics activities for the cs classroom," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, (New York, NY, USA), p. 940–945, Association for Computing Machinery, 2018.

[22] S. J. Dobesh, T. Miller, P. Newman, Y. Liu, and Y. N. Elglaly, "Towards machine learning fairness education in a natural language processing course," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2023, (New York, NY, USA), p. 312–318, Association for Computing Machinery, 2023.

[23] M. Usher and M. Barak, ""Unpacking the role of AI ethics online education for science and engineering students"," *IJ STEM Ed 11*, vol. 35, 2024.

[24] H. Shen, W. H. Deng, A. Chattopadhyay, Z. S. Wu, X. Wang, and H. Zhu, "Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 850–861, 2021.

[25] B. Walsh, S. Ali, F. Castro, K. Desportes, D. DiPaola, I. Lee, W. Payne, S. Sieke, and H. Zhang, "Making Art with and about Artificial Intelligence: Three Approaches to Teaching AI and AI Ethics to Middle and High School Students," in *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2*, pp. 1203–1203, 2022.

[26] S. K. Burriss, N. Hutchins, Z. Conley, M. M. Deweese, Y. J. Doe, A. Eeds, A. Villanueva, H. Ziegler, and K. Oliver, "Redesigning an AI bill of rights with/for young people: Principles for exploring AI ethics with middle and high school students," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100317, 2024.

[27] T. W. House, "Blueprint for an AI Bill of Rights," 10 2022.

[28] A. Padiyath, "A realist review of undergraduate student attitudes towards ethical interventions in technical computing courses," *ACM Trans. Comput. Educ.*, vol. 24, Apr. 2024.

[29] W. Revelle, "psych: Procedures for psychological, psychometric, and personality research," 01 2020.

[30] G. Baker-Brown, E. J. Ballard, S. Bluck, B. de Vries, P. Suedfeld, and P. E. Tetlock, *The conceptual/integrative complexity scoring manual*, p. 401–418. Cambridge University Press, 1992.