Teaching AI Ethics to Engineering Students: Reflections on Syllabus Design and Teaching Methods

Long paper

Lauri Tuovinen[1][0000-0002-7916-0255] and Anna Rohunen[0000-0002-4896-7056]

University of Oulu, Oulu, Finland
[1]lauri.tuovinen@oulu.fi

Abstract. The importance of ethics in artificial intelligence is increasing, and this must be reflected in the contents of computer engineering curricula, since the researchers and engineers who develop artificial intelligence technologies and applications play a key part in anticipating and mitigating their harmful effects. However, there are still many open questions concerning what should be taught and how. In this paper we suggest an approach to building a syllabus for a course in ethics of artificial intelligence, make some observations concerning effective teaching methods and discuss some particular challenges that we have encountered. These are based on the pilot implementation of a new course that aimed to give engineering students a comprehensive overview of the ethical and legislative aspects of artificial intelligence, covering both knowledge of issues that the students should be aware of and skills that they will need in order to competently deal with those issues in their work. The course was well received by the students, but also criticized for its high workload. Substantial difficulties were experienced in trying to inspire the students to engage in discussions and debates among themselves, which may limit the effectiveness of the course in building the students' ethical argumentation skills unless a satisfactory solution is found. Several promising ideas for future development of our teaching practices can be found in the literature.

Keywords: artificial intelligence, ethics education, course syllabi, teaching practices

1 Introduction

Applications of artificial intelligence (AI) involve many ethical challenges. The increasing autonomy of AI-based systems to make decisions that may have harmful consequences to living beings raises questions concerning the safety, transparency and accountability of such systems. Individuals' rights to privacy and self-determination are threatened by the collection of vast amounts of personal data to be analyzed using AI algorithms for purposes such as surveillance and psychological manipulation. There are various efforts to regulate the use of AI either currently underway or already

CEUR-WS.org/Vol-3069/FP_02.pdf

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

implemented, but given the rate at which AI technology is advancing, it seems doubtful that policymaking alone will be enough to curb morally questionable uses of AI effectively. At the forefront of this progress are the AI researchers and engineers, so as a complementary approach, we should aim to foster a strong sense of professional ethics among this community. This, in turn, means that ethics education should be included in AI engineering curricula at higher education institutions.

It is much less obvious, however, what exactly should be taught and how. For example, if we consider autonomous cars, there are some things that can be imparted to the students as facts, such as the safety record of these vehicles so far and the current legislation governing their development and deployment, but arguably the majority of interesting issues are subject to ambiguity and debate. These include relevant ethical principles and their interpretation and prioritization (e.g. how an autonomous vehicle should weigh the safety of passengers against that of other people), broader societal implications (e.g. effects of the proliferation of autonomous vehicles on the transportation sector), and implications of possible game-changing future developments (e.g. artificial general intelligence). A comprehensive syllabus thus needs to strike a balance between technical and philosophical topics, as well as between practical topics that the students can apply immediately and more theoretical ones that enable them to keep up with future developments.

Finding such a balance is a considerable challenge, especially when the time devoted to ethics in the curriculum is limited. Furthermore, the non-technical nature of much of the subject matter must be reflected in teaching and evaluation methods, which may require both students and teachers to adopt unfamiliar ways of thinking, since courses designed to build engineering skills typically do not expose them to philosophical concepts or methods. In this paper we examine the problem of teaching AI ethics to university-level engineering students based on the pilot implementation of a new course in the spring term of 2021. The course was offered at the University of Oulu, Finland, to any interested students as a non-mandatory part of a computer engineering curriculum, with emphasis on practical knowledge and skills that the students can use to identify and address ethical issues likely to be relevant to them in the present or the near future, but also some coverage of more theoretical topics.

In the remainder of the paper, we first present a review of related work in Section 2, focusing on surveys and reports of existing courses dealing with AI ethics. In Section 3 we give an overview of the objectives, topics and implementation of our own AI ethics course. In Section 4 we propose a framework for categorizing study topics and an approach to syllabus design based on the framework. In Section 5 we offer some reflections on teaching methods and practices, based on what we learned from the pilot implementation. In Section 6 we look at feedback received from students who took the course, and in Section 7 we present a discussion and conclusions.

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

2 Related work

Fiesler et al. (2020) have reviewed general courses on technology ethics in computer science education, with a subset of AI-specific courses. In their study, they have analyzed 115 syllabi with respect to course content and learning outcomes. Of these courses, 10 were AI-specific, while 55 included some content for the topic *AI and algorithms* and 27 for *AI and robots*. The authors also identified some course topics closely related to AI, such as *privacy and surveillance* with 61 courses. Their analysis results show a high variability across technology ethics courses with respect to their content, which they deem unsurprising considering the lack of standards, as well as the disciplinary breadth of the syllabi reviewed. However, they highlight the fact that the observed variability has potential to enable computing ethics educators to learn from each other and to finally begin to create norms around the learning outcomes. Despite the variability, some patterns in the syllabi reveal some critical topics, such as privacy, algorithms, inequality and justice.

Building on this work and another previous study (Saltz et al., 2019), Garrett et al. (2020) have compiled a new dataset of established syllabi, specifically focusing on AI ethics courses. They have analyzed 51 courses with AI ethics content in all, including 31 standalone AI ethics courses and 20 technical AI or ML courses with AI ethics topics integrated into them. With respect to standalone AI ethics courses, they have identified the following topics categories: bias, automation and robots, law and policy, consequences of algorithms, philosophy/morality, privacy, future of AI, history of AI. When it comes to technical AI and ML courses with AI ethics topics, the most common topics identified were bias, fairness and privacy.

Garrett et al. (2020) further identified some common teaching practices. For example, the available reading lists for standalone AI ethics courses showed that the majority of these courses included news articles as reading assignments; therefore, it seems that incorporating current events into the course material is a common way to illustrate consequences of AI usage. In some technical courses, the non-technical content focused on societal considerations or “technology for social good”. However, including ethical or social implications in technical courses may be inhibited by the fact that there is too much material to cover. Based on these two reviews by Fiesler et al. and Garrett et al., it seems that AI ethics teaching in higher education has often been organized as standalone courses so far, but in parallel with these, AI ethics topics are also increasingly being integrated into technical AI courses.

Burton et al. (2018) have developed a course where they have employed science fiction as a tool to teach AI ethics. Through this approach, they aim to move from an *authority-based view to knowledge* (typical of fields with a strong practical component and an established body of knowledge) to equipping students with skills to cope also with the unforeseen ethical issues in their future work in technology development and deployment. During the course, the students analyze science fiction stories and brief articles using ethical theories as both evaluative and descriptive tools to recognize problems and consider possible solutions from multiple perspectives. The authors suggest that their course assignments help develop capacity for attention and critical thought in a manner that serves the students in their professional lives, enabling them

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

to identify potential ethical risks related to a given technology or model, as well as to articulate their arguments about a given ethical approach and see past incomplete or specious defenses of potentially unethical projects.

Henderson (2019) has designed and implemented a standalone module on data ethics and privacy, aiming to raise students' awareness of current debates in computer science and to teach how to address these issues in practice. The module includes interdisciplinary content from ethics, law and computer science, and covers the following learning outcomes: be able to understand various conceptions of ethics and privacy; be able to critically analyze research literature at the intersection of computer science, philosophy and the law; be able to understand the effect of, and the source of, bias or discrimination in a data-intensive system; understand the need for, and optionally be able to carry out ethical, social, or privacy assessment of data-intensive projects. The module was delivered as seminar sessions, with the aim to allow deeper discussions compared to shorter lectures. The course assessment was based on an essay, peer instruction with essay-style questions on the weekly topics by the students themselves, and a data protection impact assessment task.

Wilk (2019) proposes content and teaching strategies for a new standalone course "Computers, Ethics, Law, and Public Policy" that aims to increase computer science students' ethical and legal awareness, as well as to promote critical thinking and skills needed in decision-making in their future work regarding ethical issues. He presents ethical, legal and public policy issues relevant to building and using intelligent systems. These include, for example, ethical and legal problems of algorithmic decision-making, autonomous systems, social media, fake news, journalism, privacy, and big data. The author also suggests the specific topics to be taught in the course and proposes teaching strategies supporting the course objectives; these include, for example, discussing ethical dilemmas and how to make ethical decisions, seminars with defender and opponent roles, balancing theory and practice through analyzing case studies based on ethical theories, and utilization of decision-making methodologies.

Slavkovik (2020) has designed and implemented a standalone course derived from an existing course, "Research Topics in Artificial Intelligence", where her goal was to give an overview of the core issues in AI ethics in such a way that it motivates the students to pursue further learning in this area. The course also aimed to familiarize the students with the research topic of machine ethics, as well as some of the research methods and practices in the AI field. The course topics included machine ethics, explainable AI, fair-accountable-transparent AI, and responsible AI. Learning outcomes of the course included, for example, identification of the basic ethical problems related to AI systems, understanding of the premises of the core moral theories, ability to appraise the ethical aspects of AI problems, and insight into the research process in machine ethics. Through assessment of the original course methodology, the author concluded that there was room for implementing more learning by teaching, learning by reflection, and learning by example. Scientific articles were used as learning and discussion material. The students were evaluated through an oral exam and a group project with intermediary assignments.

Fink (2018) has investigated how to find a balance in covering the theory and practice as well as the philosophy and ethics in an introductory computer science AI

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

course. She presents three examples of assignments that can be utilized to teach concepts of technical, philosophical and ethical issues related to intelligent agents in this type of courses. In the ethical assignment, students explore Isaac Asimov’s Three Laws of Robotics and the ethical implications of AI through the movie “I, Robot”. The assignment focuses on this movie as its theme revolves around Asimov’s laws and what can happen when they are applied strictly with logic and not considering standard human sensitivities. Finally, the students need to address the question of the role of pure logic vs. emotion in ethical behavior, which the author states is a key theme in the success or failure of an intelligent agent to truly make decisions that are acceptable from the human perspective.

Williams et al. (2020) have designed and implemented an experimental ethics-based curricular module for an undergraduate course, “Robot Ethics”. This module aims to teach usage of human-subjects research methods to investigate potential ethical concerns arising in human-robot interaction by engaging students in real experimental ethics research. The students participate in robot ethics research as experimenters, through which they simultaneously learn methodological approaches to experimental robot ethics and use these methods to engage with key theoretical concepts. There were three interdisciplinary learning objectives in this course: *normative influence of technology* with the aim of understanding how technologies may affect human behaviors due to their perception as moral and social agents; *experimental ethics* with the aim of understanding how human-subject experimentation can be used to explore the ethical implications of technology; and *ethical research conduct* with the aim of understanding ethical concerns that may arise in the design and execution of experimental ethics experiments. The students’ learning was assessed through an inclass quiz on the experiment’s details related to all three learning objectives.

Furey and Martin (2018) have developed a module on ethical thinking about autonomous vehicles in an AI course. They suggest incorporating this type of introductory lesson about ethics into a one-semester AI course. Through a modular approach, students have an opportunity to connect specific AI topics to the related ethical implications. In this module, students become familiar with the use of thought experiment through the trolley problem, as well as learn to understand the complex nature of ethical dilemmas. Regarding the ethics module, the course assessment is carried out as follows: the final project paper of the course requires discussion of the ethical implications of the project idea, and in the final examination there is a question assessing the students’ understanding of the trolley problem.

Shapiro et al. (2020) introduce a data ethics teaching method, *Re-Shape*, for data science education, with the aim to teach the ethical implications of data collection and use in a computing course. Through the tools and activities of the method, students collect, process and visualize their movement data. Based on these data, critical reflection and coordinated classroom activities are carried out about data, data privacy, and human-centered systems for data science. Building on the idea of *cultivating care*, students are engaged with the concept of responsibility to other and confronted with the idea that they are the “other” within systems that collect and use personal data.

Based on the AI ethics teaching examples described in this section, we have made the following remarks:

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

- The learning outcomes and objectives of the described courses and modules are relatively similar. They aim to increase students’ awareness of the ethical implications of AI usage, and to train the students to identify and address ethical issues related to the development and deployment of AI in practice.
- To achieve the learning outcomes, students are typically taught the key concepts and theories of ethics in parallel with current ethical issues related to AI usage. Furthermore, critical thinking as well as skills to discuss and appraise ethical issues are often promoted through teaching methods that support learning of this type of skills (such as article or case study analyses, or class discussions). In some courses, the students are also equipped with skills to use research methods for exploring the ethical implications of AI or trained to use practical tools to address the ethical issues.
- In higher education AI ethics teaching, a diverse set of methods and strategies have been reported. Standalone courses typically utilize assignments that comprise analysis of articles, stories or movies, discussions or seminars on the identified ethical issues, as well as consideration of how to solve the studied issues using suitable tools. AI ethics modules that can be integrated into other courses, for their part, aim to train students e.g. to use specific methods for investigating ethical issues, as well as to consider the ethical aspects of a specific application, project or activity.
- It seems that assessment methods other than traditional written exams can be employed when teaching AI ethics; this may be an approach that matches well the reported learning outcomes.

Like the majority of the AI ethics teaching examples presented above, our new course was carried out as a standalone course. Many of the existing courses and modules comprise topics and learning outcomes that are relatively similar to our ones, specifically with respect to the ethical issues of AI and the aim to provide the students with ethics skills needed in the development and deployment of AI systems in practice. Teaching methods and strategies seem to vary a great deal among the reviewed courses and modules, and an extensive set of these has been reported in the literature. Similarly to what Fiesler et al. (2020) suggested, we see this type of variability as an excellent opportunity to learn from other AI ethics educators.

Borenstein and Howard (2021) have recently presented some recommendations on how to design AI ethics teaching with the aim to foster a professional mindset for AI developers and to seriously engage the students with ethical challenges. We find these recommendations highly relevant to the selection of suitable teaching methods and strategies when teaching AI ethics to engineering students. Specifically, the authors suggest three elements to familiarize students with ethical challenges: 1) teaching the ethical design of AI algorithms, 2) incorporating fundamental concepts of data science and the ethics of data acquisition through usage of real-world datasets requiring privacy, fairness and legal issues to be addressed, and 3) offering “ethics across the curriculum” through systematic inclusion of AI ethics into the curriculum. Based on these recommendations, for instance, topical examples of misbehaving AI algorithms could be elaborated in standalone AI ethics courses, while the application of ethical principles to system design could be considered in technical AI courses.

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

3 Course description

The planning of the course, worth 5 ECTS credit points, began in 2020. The full title of the course was "AI Ethics, Privacy and Legislation". Because of the COVID-19 pandemic, the course was designed from the start to be taught remotely using the Moodle online learning platform, with lectures implemented as Zoom meetings. The learning outcomes for the course were defined as follows:

- Students will be aware of the ethical and legislative aspects and conditions that need to be considered in the design and deployment of AI applications;
- They will understand the ethical special characteristics of AI applications, compared to information technology applications in general;
- They will be able to examine existing and hypothetical AI applications from an ethical viewpoint and identify potential issues;
- They will be able to weigh the benefits of AI applications against their drawbacks also in a wider, societal context;
- They will be able to apply ethical principles in AI application design.

The course was lectured over a period of 8 weeks, with each week dedicated to one of 8 course themes. The themes are explained in Table 1. For each week of the course, two Zoom sessions were planned: a main lecture of approximately 1.5 hours and a supplementary session of approximately 45 minutes, the latter featuring a short presentation or demonstration on a specialized topic followed by discussion. Exceptions to this plan were week 5, when the supplementary session had to be cancelled due to scheduling conflicts, and week 8, when the supplementary session took the form of a tutorial for a design methodology.

The course participants were evaluated based on a series of 8 smaller assignments, one for each course theme, and a larger final assignment spanning all themes. Passing all 9 assignments was required in order to pass the course. The lecturers reviewed each submitted assignment and either accepted it as such or sent it back to the student for revisions; in either case the student would be given some verbal feedback on their work. The format of the smaller assignments varied from week to week, but as a typical example, the student would be instructed to choose an AI application and write a short essay about it, addressing questions related to the theme of the week.

For the final assignment, the students could pick one of two options. Students who attended a certain number of Zoom sessions could choose to write a lecture diary, in which they would discuss the topics covered by the lectures and supplementary sessions and reflect on what they had learned. As an alternative, students could choose to complete a final exercise, in which they would come up with an AI application concept and write a report discussing the ethical aspect of the application, guided by the 8 course themes. This second option was provided so that students who were unable to attend the lectures or preferred to work independently for other reasons could complete the course; to study the course themes, they had access to lecture slides as well as items of further reading curated by the lecturers. Each student had the opportunity to send in an

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

incomplete draft of their final assignment and receive feedback on it to help them in preparing the final submission.

Table 1. The 8 principal themes of the AI ethics course.

Week	Theme	Topics covered
1	Introduction to AI ethics	Essential concepts in ethics and AI, perspectives on AI ethics
2	Controversial AI applications	Debates and controversies involving e.g. autonomous weapons and surveillance
3	AI and data	Role of data in AI, data ethics, privacy, data ownership, data philanthropy
4	AI as decision-maker	Accountability and transparency in automated decision-making
5	AI and society	Societal changes driven/facilitated by AI, AI divides, AI literacy, good AI society
6	AI and legislation	Overview of AI regulation, active and proposed legislation relevant to AI
7	AI ethics guides	Overview of AI ethics guides, review of some major institutional guides
8	Ethical AI design	Implementing AI ethics in practice by following a design methodology

4 Syllabus design

Among the challenges of designing a balanced syllabus for an AI ethics course, there are two in particular that we wish to draw attention to. The first one of these is striking a balance between topics related to ethical reasoning and argumentation and those related to understanding AI technology and systems. Covering both areas is essential: the latter to enhance the students' awareness of AI applications and their ethical implications, the former to enhance their ability to analyze these implications and to make justified decisions regarding ethical issues when encountering them in their work. We refer to this as the **philosophy-technology axis**.

The other challenge is the rapid rate of progress in AI research and development, which raises the question of how we can ensure that at least some of the course content remains relevant to the students also in the future. For dealing with issues that are immediately relevant to today's AI practitioners, the course can offer practical tools such as ethical design guidelines and methodologies, but the further into the future we look, the less we know about what the capabilities of AI will be, and therefore the less

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

likely it is that any concrete guidelines given today will be useful in dealing with the ethical implications of those capabilities. To help the students gain some degree of preparedness, the course should cover technology topics representing visions of the future of AI, as well as philosophy topics that are more theoretical in nature but also more stable over time. This we refer to as the **practice-theory axis**.

If we consider these two axes as perpendicular dimensions, we can visualize them as shown in Fig. 1. We can thus divide possible study topics in AI ethics into four broad categories, corresponding to the four quadrants of the figure. Counterclockwise from the top left, these are as follows:

- *Timeless Foundations*: established concepts, theories and traditions in ethics;
- *Practical Guidance*: applied ethics principles and guidelines relevant to AI;
- *Here and Now*: AI applications of the present and the near future and the ethical issues associated with them;
- *Beyond the Horizon*: future potential of AI and the ethical implications of hypothetical scenarios.

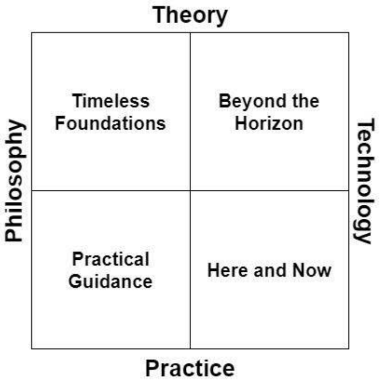

Fig. 1. Categorization of study topics in AI ethics based on two perpendicular axes.

Fig. 2 shows a selection of possible topics to be discussed on an AI ethics course, placed approximately where they are located on these two axes. The grey ellipse represents an approximation of the area from where we took the bulk of the subject matter for our course. Notably, the course syllabus was centered on the major ethical issues identified in present-day AI applications – safety, bias, accountability, explainability, privacy – and on the topics in the immediate vicinity of these, which provide essential context for the current issues and/or means for dealing with them. The least attention was devoted to the topics at the top of the figure, i.e. the most abstract and/or futuristic ones such as metaethics and superintelligence.

We argue that this represents a reasonable core syllabus for a practically oriented AI ethics course, and furthermore that this system for classifying study topics can be used as a practical framework for designing a syllabus for a course with a wider scope. Expanding the scope with respect to a given aspect of AI ethics can be thought of as stretching the grey scoping bubble; the visualization of affinities between topics is useful in ensuring that the course remains a coherent whole, since it suggests clusters

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

of closely related topics that it would make sense to include together (and possibly also to discuss together, so the framework provides some hints for planning the course structure as well). The overall shape of the bubble serves as an indicator of the balancedness of the syllabus, so if the bubble becomes strongly elongated in some direction, this suggests that reviewing the planned syllabus and possibly including some additional topics for the sake of balance should be considered.

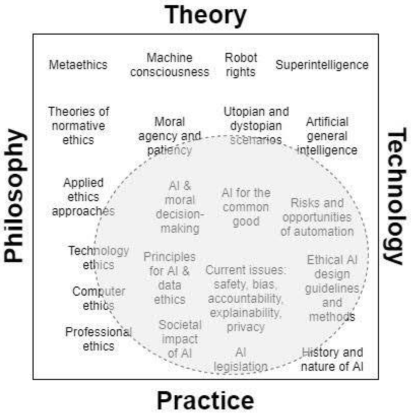
Theory
Philosophy
Technology
Practice
Metaethics
Machine consciousness
Robot rights
Superintelligence
Theories of normative ethics
Moral agency and patiency
Utopian and dystopian scenarios
Artificial general intelligence
Applied ethics approaches
AI & moral decision-making
AI for the common good
Risks and opportunities of automation
Technology ethics
Principles for AI & data ethics
Current issues: safety, bias, accountability, explainability, privacy
Ethical AI design guidelines, and methods
Computer ethics
Professional ethics
Societal impact of AI
AI legislation
History and nature of AI

Fig. 2. Placement of topics on the philosophy-technology and practice-theory axes.

5 Teaching methods

A central challenge that we faced in the planning and execution of the course was presenting the subject matter to engineering students in such a way that it engages them to think about AI more philosophically than their other courses require them to. Toward this end, each week of teaching would follow the same general pattern, where the lecture would introduce the students to the theme of the week from a mostly theoretical perspective. The role of the supplementary session would then be to have the students discuss their thoughts with their peers, and the role of the assignment would be to have them apply what they had learned.

Real-world examples of AI applications were used in the Zoom sessions to illustrate theoretical concepts, and in many of the assignments to provide the students with a context where they can demonstrate their understanding of the subject matter. Perhaps predictably, when instructed to choose an application to analyze, the students tended to choose ones that are close to their own everyday lives; for example, services offered by Facebook or Google were popular choices. The evaluation of the assignments was based mainly on the students' ability to build and defend arguments using the concepts and theories discussed in the Zoom sessions, although some of them also included a fact-finding element. Furthermore, several assignments invited the students to show original thought by proposing ideas for addressing ethical issues associated with a given application or technology.

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

Another category of AI applications that engaged the students' interest were those involving palpable risk to core values such as health, safety or liberty. This is suggested by the popularity of such applications (e.g. medical AI systems, autonomous vehicles/weapons, predictive policing applications) as assignment subjects, and also by active audience participation during the week 2 lecture on controversial AI applications. The first recorded case of a pedestrian killed by an autonomous car was used already in the first lecture as an introductory example, since it provides an effective way to illustrate a number of key points:

- A malfunction in an AI system may have severe harmful consequences, including loss of life.
- When this happens, it may be hard to ascertain why it happened and who is morally or legally responsible.
- Even a hypothetical perfect AI system may encounter a situation where it has to make a value-based choice between multiple bad options.
- Short-term harm may be inevitable in the pursuit of long-term good, and the justification of such harm is also debatable.

Overall, the course achieved mixed success at best in inspiring discourse and debate among the students. This represents a missed opportunity for the students to learn valuable argumentation skills through peer-to-peer interaction. We can only speculate on the reasons here; presumably, the teaching methods employed were less than optimal for this purpose, but on the other hand, it may also be that many of the students (who were computer engineering majors) were not predisposed to working like this. Furthermore, it seems plausible that a physical classroom where the participants can better connect with one another would have been a more suitable environment for this. This was particularly evident in the final supplementary session, where the participants were given a tutorial for a design methodology based on a set of cards, designed to be printed out and manipulated as physical objects. Since this was not an option, the tutorial was implemented using an online whiteboard application, with the result that the students participated mainly by adding their own ideas to the whiteboard without interacting with one another on Zoom while doing it.

Online lecturing did have the unanticipated advantage that the students could use Zoom chat to ask questions and exchange thoughts even while the lecturer is speaking. We also set up a message board on Moodle and encouraged the students to use it for discussions, aiming to engage also those students who feel less comfortable expressing their views live on Zoom, but this proved unsuccessful. Another benefit of remote teaching was that this made it convenient to include lectures and presentations by visiting experts in the course program. Visitors were featured in two of the main lectures and three of the supplementary sessions; since all of these were planned from the start to be implemented as Zoom meetings, the visitors could deliver their presentations from where they are normally based and no special arrangements were required, making physical distance a non-factor in deciding which external experts to invite as visitors. Besides contributing their expertise to the course, the visitors enabled the principal lecturers to focus their efforts on course themes closer to their own respective areas of expertise, resulting in a higher overall standard of teaching quality than would otherwise have been achievable.

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

6 Student feedback

Feedback on the course was solicited from the participating students using two anonymous surveys. A standard feedback form, generated automatically for every course at the university, was filled in and submitted by 7 course participants. The form included 11 statements to be rated on a Likert scale from 1 (totally disagree) to 5 (totally agree). The results are summarized in Table 2. The students were also requested to evaluate their workload during the course, bearing in mind that 1 ECTS credit is equivalent to 27 hours of student work. The answer options and corresponding integer values were no workload (2), light workload (1), suitable workload (0), high workload (1) and very high workload (2); the answers were split between suitable workload and very high workload for an average of 1.14.

Table 2. 11 statements about the course, rated by students on a Likert scale.

Statement	Avg
I achieved the course learning outcomes and course objectives.	3.86
Course content helped me to achieve learning outcomes.	3.57
The course content supported my progression towards expertise in my field.	4.00
Teaching methods supported learning and helped me to achieve learning outcomes.	3.71
Course material supported my learning.	3.57
Instructions to the course tasks were clear.	3.86
There was enough support and guidance in the course.	3.57
Assessment methods and criteria supported my learning.	4.14
There was enough time to complete the tasks in the course.	4.14
I have tried to advance my own learning outside the lectures (e.g. reading the lecture material, reading some literature connected to the topic or searching more information).	4.43
I have tried to advance my own learning during the lectures by discussing the topic with other students, by asking questions from the lecturer, by starting discussion in the whole group or questioning the teaching.	3.71

In addition to these, the standard form had open-ended questions on good practices, things to develop and other feedback. Notably, several of the answers to these questions included complaints about the workload of the course assignments. There were also individual criticisms concerning the course timetable and the supplementary session concept. Positive feedback was received on lectures, study materials, weekly assignments and guest lecturers.

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

A supplementary feedback form was designed to capture more detailed information on the students' expectations and learning outcomes. This form was available in the course workspace on Moodle and was filled in and submitted by 7 course participants. Concerning expectations, the form included the following questions:

- What were your reasons for signing up for the course?
- What were your expectations for the course with respect to your personal and professional development?
- Would you agree that the course fulfilled your aforementioned expectations?

The first two of these were open-ended questions. Among the answers were some generic motivations such as finding the subject interesting and needing the credits to complete one's studies, but several answers also indicated that the student was expecting knowledge of AI ethics to be an asset in jobs involving AI. The third question was to be answered on a Likert scale from 1 (strongly disagree) to 5 (strongly agree); the answers ranged from 3 to 5 for an average of 4.29. Optionally, the students could specify their answers to the third question; here one student indicated that they would have liked to learn more about the fundamentals of ethics, while another remarked that the exercises required a lot of effort and more writing than the student was used to, but that they were also good practice. Better visualization of lecture materials was also requested, and numeric grading instead of pass/fail was suggested as an incentive to put more effort into the assignments.

Concerning learning outcomes, the form included four Likert-scale questions corresponding to the four basic categories of study topics identified in Section 4. The theory here was that a balanced syllabus would contribute to the students' *awareness* of ethical issues in AI – both those that are relevant now and those that are still in the future – as well as their *confidence* in their ability to deal with these issues in their work. The results are summarized in Table 3.

Table 3. 4 statements about course outcomes, rated by students on a Likert scale.

Statement	Avg
The course increased my awareness of ethical issues that are relevant to AI applications today.	4.71
The course increased my confidence in my ability to competently address these current issues in my work, if and when I encounter them.	4.14
The course increased my awareness of ethical issues that may arise in the future, given the development potential of AI technology.	4.43
The course increased my confidence in my ability to competently address these future issues in my work, if and when they arise.	4.29

7 Discussion and conclusions

In this paper we discussed the teaching of AI ethics to engineering students, based on experience of lecturing a pilot course in the spring semester of 2021. We faced several

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

challenges in designing a comprehensive and balanced syllabus for the course, as well as in the selection of effective teaching methods. The positive feedback received from participants suggests that the course was at least a moderate success, but since the number of students who took the course was small and only a fraction of them answered the feedback surveys, the results mut be interpreted with caution.

The aspect of the course that received the most criticism from the students was the workload. The most obvious candidate for an explanation is that the effort required to complete the assignments was underestimated when planning the course, but there are other possible contributing factors. For example, if the format of the assignments was such that it took the students out of their comfort zone – as engineering students they would be more familiar with e.g. programming assignments than essay writing – then this may have affected the students' perception of the course workload. Another possible factor is that both the individual assignments and the course as a whole were graded on a simple pass/fail grading scale, so there were no rewards other than words of praise from the evaluators to be earned by surpassing the minimum requirements (as one student explicitly pointed out in their feedback). This is perhaps the most important issue to be addressed when planning future iterations of the course.

As another student noted in their feedback, commenting on the workload, the course format required a lot of effort from the teachers as well, which is an important point since it raises a question concerning the scalability of the course. Out of an original 30 students who signed up for the course, roughly half completed all the assignments – the high dropout rate perhaps another indicator that the students found the workload higher than expected – which kept the effort required to review the submissions manageable, but had all 30 completed the course, the allocated teaching resources would have been stretched to their limits. The issue here is that designing an alternative to the current assignment format that scales up for a significantly larger number of students while still effectively building and testing their ethical argumentation skills is far from trivial. Getting the students engaged in debates among themselves would both develop these skills and reduce the burden on the teachers, so again the crucial question is how to effectively facilitate such debates, given that there may be many students who are not comfortable with this type of work. This is another major aspect of the course that will need to be addressed in future development.

Among the practices reported in the studies reviewed in Section 2, there are some that seem particularly useful and promising for the future development of our teaching. For example, as there is still a lack of established textbooks, we should maintain and regularly update our current reading list and make sure that it includes both scientific papers and news articles (the latter to illustrate the real-world consequences of AI use). Due to the rapid technological development and changes expected in the field, the need to update the list should be reviewed regularly. As engineering is a typical field with a strong authority-based view to knowledge, it could be fruitful to think even more about how to move from this type of thinking to providing our students with skills to deal with unforeseen ethical issues in their future work tasks. Some of the proposed methods in the reviewed studies could fulfill this need well and would fit our course implementation, such as seminars with formal defender and opponent roles or case study analyses. Finally, we may also consider whether we should extend AI ethics

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

teaching to reach a larger portion of the students, as well as how to make the connection between technical AI topics and their ethical implications clearer; in addition to our current standalone course, AI ethics modules or assignments could be integrated into technical AI courses starting from the early stages of engineering study programs.

References

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, 1(1), 61–65. https://doi.org/10.1007/s43681-020-00002-7

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

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

Fink, P. (2018). *Addressing the Technical, Philosophical, and Ethical Issues of Artificial Intelligence Through Active Learning Class Assignments*. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), Article 1. https://ojs.aaai.org/index.php/AAAI/article/view/11401

Furey, H., & Martin, F. (2019). AI education matters: A modular approach to AI ethics education. *AI Matters*, 4(4), 13-15. https://doi.org/10.1145/3299758.3299764

Garrett, N., Beard, N., & Fiesler, C. (2020*). More Than "If Time Allows": The Role of Ethics in AI Education.* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 272–278. https://doi.org/10.1145/3375627.3375868

Henderson, T. (2019). *Teaching Data Ethics: We're going to ethics the heck out of this*. Proceedings of the 3rd Conference on Computing Education Practice, 1-4. https://doi.org/10.1145/3294016.3294017

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., & Beard, N. (2019). Integrating Ethics within Machine Learning Courses. *ACM Transactions on Computing Education*, 19(4), 32:1-32:26. https://doi.org/10.1145/3341164

Shapiro, B. R., Meng, A., O'Donnell, C., Lou, C., Zhao, E., Dankwa, B., & Hostetler, A. (2020). *Re-Shape: A Method to Teach Data Ethics for Data Science Education.* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3313831.3376251

Slavkovik, M. (2020). *Teaching AI Ethics: Observations and Challenges*. Norsk IKT-Konferanse for Forskning Og Utdanning, 4, Article 4. https://ojs.bibsys.no/index.php/NIK/article/view/815

Wilk, A. (2019). *Teaching AI, Ethics, Law and Policy*. ArXiv:1904.12470 [Cs]. http://arxiv.org/abs/1904.12470

Williams, T., Zhu, Q., & Grollman, D. (2020). *An Experimental Ethics Approach to Robot Ethics Education.* Proceedings of the AAAI Conference on Artificial Intelligence, 34(09), 13428-13435. https://doi.org/10.1609/aaai.v34i09.7067

Copyright © 2021 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)