Teaching AI Ethics: Observations and Challenges

Marija Slavkovik

Department of Information Science and Media Studies, University of Bergen, Norway

marija.slavkovik@uib.no

May 2020

Abstract

This report summarises the experience in teaching Artificial Intelligence (AI) Ethics as an elective masters level course at the University of Bergen. The goal of the summary is twofold: 1) to draw lessons for teaching this in-high demand very new discipline; 2) to serve as a basis in developing a bachelor level AI Ethics course for students of artificial intelligence. AI Ethics as a topic is particularly challenging to teach as the discipline itself is very new and no textbooks have been established. The added challenge is introducing methodologies and skills from humanity and social sciences to students of computational and information sciences.

1 Introduction

On January 14, 2020, the Norwegian National strategy on Artificial Intelligence was presented by the Ministry of Local Government and Modernisation (2020). The Norwegian national strategy follows a trend in national, international and professional guidelines and recommendations on how artificial intelligence (AI) systems should be developed and used[1]. The national strategy specifically stipulates:

> "Algorithms can be controlled by facilitating access or audit, but it is more appropriate for developers as well as users to build privacy and ethical considerations into systems from the outset."(Ministry of Local Government and Modernisation, 2020, pg. 60).

At present, courses in artificial intelligence and computer programming are typically offered to students of "technical" studies such as computer science, informatics, engineering and information science, whereas subjects which develop skills that would help one to recognise and understand issues of "privacy and ethical considerations" are typically offered to students in humanities and social sciences. In other words, in Norway, and generally in the world, it is not part of the higher education tradition to train either "developers" or "users" to build "privacy and ethical considerations into systems" (O'Neil, 2017).

The Department of Information Science and Media Studies (Infomedia) at the Faculty for Social Sciences at the University of Bergen offers the course Research Topics in Artificial Intelligence (code INFO381) as a 15 point masters course. The course is intended to be tailored by the lecturer to present a specific topic of research in artificial intelligence from that lecturer's area of expertise. I have been given the opportunity to teach INFO381 in the spring semester of 2017 and again in 2020. Having an active research interest in AI Ethics, I have designed a INFO381 variant to give masters students an introduction the state of the art and main challenges in AI Ethics research.

[1] See for example Algorithmic Watch (2020) for an exhaustive list of guidelines.

AI Ethics has gradually emerged as a field of artificial intelligence in the past fifteen years. The field of AI Ethics is an active research field, however it is not sufficiently established to offer an indisputable curriculum for teaching it. At present, there is an increasing interest to offer AI courses to future bachelors of informatics, information science and computer - if we are to follow the national guidelines on AI, the skills in AI have to be complemented with awareness of responsible use.

> The main motivation for this report is the question: How to design a bachelor level course on AI Ethics for the curriculum of future AI developers, researchers and data scientists?

Specifically, I investigate how to leverage my experience in teaching AI ethics at the masters level to design a bachelor level course on the same topic. Apart from the acute societal need for such a course, the Infomedia Department is in the process of offering a new bachelor program in AI. AI Ethics is a new course currently in development that will be offered as part of this program.

This document is structured as follows. In Section 2 I detail the topic of AI, AI ethics, and I give an overview on the state of education in AI ethics in the world. In Section 3, I outline how AI ethics has been thought in INFO381 in 2017 and outline the lessons which were used for planing the same course in 2020. In Section 4, I detail the observations from teaching INFO381 in 2020, the lessons learned, and how those lessons can be leveraged into a plan for an bachelor level AI ethics course. In Section 5, I outline the bachelor AI Ethics course. Lastly, in Section 6, I summarise this report and outline the main insights.

2 Background

2.1 AI: what it is and how it changed

Artificial Intelligence (AI) is an interdisciplinary research and application field rooted in computer science, mathematics, and information science. It traditionally also involves research in philosophy, cognitive science, social science, economy and law. AI is concerned with issues of creating intelligent behaving computational agents and understanding what makes a computational agent behaviour intelligent (Russell und Norvig, 2015). AI can also been seen as the science and practice of automating cognition (Bellman, 1978).

Artificial intelligence has been established as a research field in 1956 (Moor, 2006a). Prominent applications of methods developed in artificial intelligence have started to appear already in the 1980s. Like all computing at the time, software and autonomous devices were seen as tools which are developed by professionals to be used by professionals. Thus all users and developers were fully educated into the abilities of the AI system they worked with. Furthermore, the contexts in which a given system was used were known and controlled. In other words, both software and hardware existed in a so called "working envelope" segregated from society with precisely defined area of operation and human access control.

Since the new millennium, computing is no longer something that happens on computers in front of us and computers are no longer only a professional work tool. Computing and computers are now ambient and ubiquitous in the sense that a person is not necessarily aware they are interacting with a computer or that they contribute to computing. Supported by an increase of computing power and availability of structured data, branches of AI research become applicable into many domains in solutions that now cater not only to the train professional but to the general public. Specifically the sub-areas of machine learning, vision, and natural language processing require a lot of untrained cognitive labour to create examples used to "train" algorithms to recognise faces in images, translate from one language to another, transcribe speech to text, dynamically price travel tickets, etc.

In the last decade we see a reinforcement in the general trend of computation leaving the "working envelope". There is now unconstrained interaction between software and people. Programming is done not only by formally trained professionals, but also by self-taught people and enthusiasts. AI systems

now replace and augment human cognitive activities such as decision-making in a variety of domains. With this trend, concerns arise on how to ensure that the AI systems developed are aligned with the values of its users, developers and the society in which they are deployed (Dignum, 2019).

2.2 AI Ethics

AI ethics is the common reference to a collection of sub-fields in AI developed to respond to the issues of how to manage the moral, personal and societal impact of replacing people tasks and roles with AI powered computing. AI Ethics comprises of four main research sub-disciplines: fair-accountable-transparent AI (FAccT), explainable AI (XAI), responsible AI, and machine ethics (also called artificial morality). I introduce each briefly.

Bovens (2007) defines accountability as:

> "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences." (Bovens, 2007, p. 447)

In AI ethics we focus on *algorithmic accountability* which specifies an accountability relationship in which the actor gives account for an algorithm which they may or may not have created (Wieringa, 2020). An algorithm is a model of software that describes instruction for a computer. The role of the author can be filled by anyone involved in the algorithm's creation and deployment. The role of the forum can be filled by different stakeholders that are affected by the algorithm's operation, such as: users, government supervisory bodies, other companies etc.

Fairness motivates algorithmic accountability. The field is concerned with ensuring that an algorithm is equally efficient for all users and that the continuous use of an algorithm, particularly in a decision-making role, does not introduce a source of discrimination in society (Mehrabi u. a., 2019). The complementary concern of the field are the various types of fairness that one can define or be interested in protecting, and also how do these different fairness requirements interact with each-other and with the computational efficiency of the algorithm itself.

Transparency refers to the algorithm that is the subject of accountability and the problem of making sure that the forum is able to inspect and understand the algorithm itself (Weller, 2017). This field is also concerned with establishing a framework for defining different types of transparency but also with the question of: how to write algorithms that can be automatically audited?

Explainable AI (XAI) refers to the account artefact in the algorithmic accountability relationship (Gunning und Aha, 2019). One aspect of this sub-field is understanding what makes an account an explanation for a particular actor and forum (Miller, 2019b). The other aspect of explainability is developing algorithms that can generate explanations for their operations or outputs, and developing algorithms that can interpret the operations or outputs of complex algorithms (Arya u. a., 2019).

Responsible AI is concerned with ensuring that the forum has sufficient power over the actor in the algorithmic accountability relationship. Unlike products that you can touch, software and data can exist and be used in more than one country at the same time. This makes legal regulation of AI very difficult. The filed has numerous sub-interests that include: establishing a professional code of ethical conduct for AI researchers, developers and practitioners; developing strategies for assessing the ethical impact and value alignment of a particular AI application. A big part of responsible AI is overseeing the development of AI guidelines and ensuring that they are meaningful and actionable (Jobin u. a., 2019; Hagendorff, 2019).

Machine ethics (Moor, 2006b), or artificial morality (Wallach und Allen, 2008), unlike the other AI Ethics sub-fields studies how to program moral behaviour in machines and software. The questions studied in this field include, but are not limited to: what moral theory to implement, should moral behaviour be learned or hard-coded, what characterises ethically sensitive decisions and contexts, how to decide what moral values should a machine enforce.

All of these different sub-fields of AI Ethics have been developing at different rates. The communities of researchers and practitioners is not well connected spanning many scientific disciplines. However, the interests of the sub-fields are clearly related and at points overlap, therefore an introductory level course in AI Ethics should include all of them.

2.3 Global education in AI Ethics

AI as a subject is typically only thought at an university undergraduate level. It is now certain that AI applications are a socially disruptive technology. On one hand the AI applications change aspects of society and it is in our own interest to know how. For example, we may want to be aware that we are interacting with an AI customer service decision-maker not a human and adjust our communication strategy accordingly. On the other hand, AI applications directly displace jobs changing the range of required skills that make one employable. Education is one of the tools we have to ensure that we understand what we are disrupting with a particular AI application, but also to ensure that we enable people to do the jobs of the future. In response to this reality some governments are going so far as to include AI education already in primary schools (Chan, 2019). Universities in the Western World have been criticized that they are "asleep at the wheel, leaving the responsibility for this education to well-paid lobbyists and employees who've abandoned the academy" (O'Neil, 2017).

In response to the O'Neil (2017) article, Fiesler u. a. (2020) studied the integration of ethics into computer science university curriculums. The authors did a qualitative analysis of 115 syllabi, from a total of 202 such identified AI Ethics courses in university technology curricula. They considered the courses content and learning outcomes. The courses considered in the study were identified through a crowdsourced collection of "tech ethics" syllabi (Fiesler, 2018). The authors also looked into who teaches the courses.

Arguments have been put forward that defend the position that ethics philosophers should teach AI ethics course (Johnson, 1994), but also to defend the position that this is a task for computer science lecturers (Grosz u. a., 2019; Fiesler u. a., 2020). Fiesler u. a. (2020) found that the majority of the courses were thought at computer science and information science departments and were given by lecturers from these departments. Most courses were thought at the under-graduate level (107/115), 74 of the courses were thought at the graduate level, and 19 cross-listed for both. The courses were given both at the beginning and at the last years of studies.

The topics covered in the studied "tech ethics" courses included: Law & policy (66), Privacy & surveillance (61), Philosophy (61), Inequality, justice & human rights (59), AI & algorithms (55), Social & environmental impact (50) Civic responsibility & misinformation (32), AI & robots (27), Business & economics (27), Professional ethics (25), Work & labor (23), Design (20), Cybersecurity (19), Research ethics (16) and Medical/health (12). The goals and learning outcomes of the studied courses included: Critique, Spot issues, Make arguments, Improve communication, See multiple perspectives, Create solutions, Consider consequences and Apply rules. Fiesler u. a. (2020) pointed out to a great variability in content across courses and disciplinary breadth.

An overview of the results of Fiesler u. a. (2020) and also the raw data suggests that none of the reported courses includes an all-encompassing overview of the AI-Ethics disciplines. Furthermore no courses teach machine ethics. The reported course topics are organised around specific societal and moral values rather than AI ethics disciplines. In 2017, INFO381 was one of the first courses in machine ethics in the world [2].

[2] I am aware of the syllabus of an undergraduate level course at the University of Saarbrücken created by Kevin Baum and Holger Hermanns.

3 INFO381 - 2017

3.1 Execution

When I constructed the INFO381 course in 2017, the field of AI Ethics was not as advanced as it is today. The topic of the course was limited on machine ethics as this sub-discipline had existed in computer science AI since at least 2006. Since 2017 the fields of FAccT, XAI and Responsible AI have been established with well visible international publishing venues, but also with considerable media and big tech company attention. I first describe the of INFO381-2017 course content, goals, learning outcomes and evaluation, before discussing the results and observed challenges.

INFO381-2017 was organised in eight, six hour long sessions. This is the standard organisation form for the Masters in Information Science courses at my department. Each session was structured as a sequence of lectures, discussions and group work. The list of sessions and material used is given in Table 1. The curriculum was left relatively open for adapting to the learning and reading pace of the students.

1. Introduction to machine ethics	Chapters 1,26,27 from Vaughn (2014), Moor (2006b), Wallach u. a. (2008), Bonnefon u. a. (2016), Winfield u. a. (2014)
2 Introduction to moral philosophy	Chapters 4,6,7 and 12 from Vaughn (2014)
3.Bottom-up approach with an introduction to machine learning	Anderson und Anderson (2008), Chapters 18,19,20,21 from Russell und Norvig (2015)
4. Top-down approach with an introduction to agent verification	Dennis u. a. (2016), Chapter 7 from Russell und Norvig (2015)
5.Computers do what you ask them to, not what you want them to. Limitations of top-down and bottom-up approaches	Lecture Notes
6. Context and impact of machine ethics - cultural implication, military vs civilian applications, algorithms vs. embodied intelligent systems	Bolukbasi u. a. (2016), Sharkey und Sharkey (2012), Arkin (2008)
7. Deliberately making unethical machines	Pistono und Yampolskiy (2016)
8. Research methodology instructions	Lecture Notes

Table 1: Topics and materials used in INFO381-2017.

The goal of the course was to familiarise the students with the research topic of machine ethics specifically, and with some of the research methods and practices used in artificial intelligence in general. The learning outcomes for this course were as follows.

> **Knowledge.** Identify and describe the basic principles of moral philosophy, interpret, explain and extend the need for, and challenges of, automating moral reasoning. Experience the entire process of research in machine ethics from the inception of an idea, analysis of research work, refining a research question, planing and executing group work and reporting on the work in the form of a scientific report.
>
> **Skills.** Appraise ethical aspects of AI problems. Discern different moral theories and values when considering ethical impact of AI applications.
>
> **General competence.** Reading and explaining scientific articles. Research project management. Scientific reporting.

The students were evaluated through an oral exam (40% of the grade) and a group project (60% if the grade). In addition, there were two additional obligatory assignments subject to approval. The evaluation methods were designed to be in service or learning rather than grading. The exam components were maximally personalised to allow the students freedom in following their interest while learning new skills, but also to relate their own performance with the effort put in rather than compare with the performance of other students. The goal of each component was to harness the incentive that a good grade brings for pushing oneself to independently learn.

The oral exam was executed as a presentation. The students were tasked with finding and selecting an article in machine ethics, which they were then to read, presenting and answer questions about in a 25 minute presentation. The students were allowed to select any article in the scope of the discipline. The proposed articles were approved by the lecturer, to exclude position papers, short papers and papers entirely within philosophy, psychology and social science with only a casual link to artificial intelligence. The students were graded on the quality and clarity of the presentation, ability to answer questions and ability to analyse the scientific contribution in the context of the other work read throughout the course. The learning goal of this exam was to: 1) ensure that the students thoroughly read and understand a piece of scientific literature and become aware of the cognitive task and time involved; 2) experience public speaking and presentation of scientific topics and gain confidence into doing this type of activity.

The obligatory assignments were used to initialise the student group projects. In the first assignment the students were asked to identify an AI application that has an ethical impact (on a person, or society) and describe this application and perceived impact in a short one page deliverable. The goal of this assignment was to open up the views of the students into where AI applications are in our society and nudge them into questioning the ethical impact off these applications.

As a second assignment the students were asked to prepare (individually or in pairs) a research idea pitch for the rest of the class. The ideas were delivered in written form, a one page description of a research question, motivation, suggested methodology and success criteria. In defining the methodology, the students were asked to use at least one reference to a scientific article. They were also asked to describe the pitching pair skills and the required skills for completion of the project. The project ideas were pitched in the class and the class voted for the projects that they would like to execute. INFO381-2017 was attended by 25 students, 14 projects were pitched, of which 4 were selected for execution.

The pitching idea students who have won the popular vote were tasked with selecting a team and organising the execution of the project. The students worked independently developing their project ideas, however they were given supervision in terms of pointers to methodology and literature, scoping of research question and writing out the proposals. The goal of each of the project work was to experience an authentic practice of creating research.

Having the students design and choose their own project topics contributed to the students feeling that they are pursuing their interests rather than being forced into work. The students choose projects that helped them benefit from the skills they have gained so far in their studies but also they chose projects that used skills they wanted to acquire. Furthermore, because they designed and chose their own work they could choose how much work they will do, thus helping them understand hands-on how to plan work, what constitutes research progress and advancement of the state of the art. Most importantly, the students were aware that there are no ready answers or correct solutions to the problems they were engaged with, unlike project work where the task is defined by the lecturer.

The work of each project was described into a scientific report. The students used scientific articles as a template for the report. Rather than being assigned a specific typesetting format for the reports, they were asked to focus on clarity and completeness of the described work. All of the reports fell between 7000 and 10.000 words. These projects were eventually graded: A, B, B and C. The goal of the written report assignment was to get the students to experience the challenge of scientific expression. Writing what was essentially a scientific article was also meant to help them understand better the articles the were reading.

After the grading was completed the students were given the opportunity to publish their project reports. This was optional and not related to the grade. The students were to select one person from their group to reformat the report into a publication and eventually present it to the scientific venue. All groups chose to do this. Out of the four projects, three were published as articles: Pires Bjørgen u. a. (2018), Caycedo Alvarez u. a. (2017), Valland u. a. (2017). The department made funding available for the presenting student to attend the respective venue and experience that side of research as well. Going thought this final step helped the student see the grades as a learning opportunity rather than a judgement and disassociate grades from value. They also experienced the peer-review process.

3.2 Observations

One of the initial practical challenges to the designing INFO381-2017 was the lack of AI courses offered at the University of Bergen. Specifically machine learning was not being thought at the time, while being the fastest growing AI discipline and one most frequently considered to have ethical impact. This considerably limited the research options for students and curriculum. In 2018 I have developed Machine Learning as INFO284 at the Department of Information Science and Media Studies, while the Informatics Department introduced machine learning as INF283 in 2019 in addition to other machine learning courses.

Teaching a subject that is only a decade old means that there exists no textbook that can be readily used. There was not even sufficient time to establish which scientific articles are most influential in the field. Scientific articles are often not written for general audience - they assume the reader has a certain lever of familiarity with the subject domain. The masters students were not very accustomed to reading such literature and were not comfortable with the domain of artificial intelligence or philosophy. All this significantly limited the available resources for the course. The selection of articles had to be sufficiently diverse to cover the broadness of topics in machine ethics, but the articles themselves had to be sufficiently established and well regarded in the AI community.

Perhaps the most challenging aspect of INFO381-2017 was in the subject matter of moral philosophy. The focus of the training of students in technical courses is their ability to "think in an algorithmic way" to see structured data in knowledge, to transform a problem into an algorithm, break it down into smaller programs and from there identify programming instructions. The masters students of information science are not an exception.

In mathematics and computer science there is a correct solution and an incorrect solution. We "train" students to check their solutions for correctness. Ethical problems and philosophical problems however rarely have an objectively correct solutions and this state of affairs also holds in aspects of machine ethics. Handling tasks for which there is no correct solution presented a bigger challenge for the students than anticipated. Such problems need to be addressed with arguments rather than proofs. Although I designed the courses around discussion topics intended to offer practice in argumentation, the offer did not meet the need. The students were also only offered one round of feedback to their project reports and this also turned out to be insufficient.

4 INFO381 in 2020 execution and observations

4.1 Execution

The observations outlined in Section 3.2 were considered when planing INFO381-2020. The course methodology was also assessed with respect to strategies from the learning sciences Bransford u. a. (2000). While key learning strategies such as learning by doing, authentic practices, project-based learning, personalized learning, collaborative learning and immersion in a community of practice were operationalised in the course, there was a clear room for implementing more learning by teaching, learning

by reflection and learning by example.

Two critical changes impacted the planning of INFO381-2020: we were now teaching machine learning to our students and the AI Ethics field has exploded with work in XAI, FAccT and Responsible AI. The learning outcomes of the course and syllabus were adjusted in this regard to reflect the students' now existing knowledge in AI:

> **Knowledge.** Identify the basic problems studied in XAI, FAccT, Responsible AI and machine ethics. Understand the premises of the core moral theories. Interpret, explain and extend the need for, and challenges of, AI Ethics. Experience the entire process of research in machine ethics from the inception of an idea, analysis of research work, refining a research question, planing and executing group work and reporting on the work in the form of a scientific report.
>
> **Skills.** Appraise the ethical aspects of AI problems. Match a specific AI Ethics challenge to its most relevant discipline.
>
> **General competence.** Reading and explaining scientific articles. Research project management. Scientific reporting.

INFO381-2020 was again executed in eight six, hour session, however, three hours in each of the last four sessions were devoted to working with the students on their projects and reports. Furthermore, very early on the students were explicitly instructed in argumentation theory and challenged to engage in discussions on assigned reading material. Unfortunately, the Covid-19 crisis and closure has interrupted regular classes and this plan for increasing the discussion in class was not entirely executed.

As learning and discussion material again scientific articles were used. The articles were mainly given as a reading assignment before each class. Additional home activities were given for students to try out. Each session also included a workshop on various forms of scientific reporting. The list of sessions and material used is given in the following list:

1. What do we talk about when we talk about AI and Ethics
   Material: Chapters 1 and 2 from Russell und Norvig (2015)
   Discussion: Ministry of Local Government and Modernisation (2020)
   Try out: `http://moralmachine.mit.edu/`

2. Basics of moral philosophy 1/2
   Material: Vaughn (2014)
   Discussion: Turing (1950), Weizenbaum (1966), Searle (1980)
   Workshop: how to motivate a research question
   Try out: find out who Eugene Goostman is

3. Basics of moral philosophy 2/2
   Material: Vaughn (2014)
   Discussion: Rawls (1958), McIntyre (2019), Anderson (2008), Foot (1967)
   Workshop: role of related work in science articles
   Try out: 1. fix the three laws of robotics (or come up with your own) 2. represent in a representation of your choice "do not do harm"

4. Deontic logic, Top-down machine ethics
   Material: Lecture notes
   Discussion: Moor (2006b), Amodei u. a. (2016), Awad u. a. (2018), Powers (2006)
   Workshop: research question and success criteria
   Try out: How would you build a prima facie duty following implicit artificial agent? How would you decide what moral theory should be used for governing the behaviour of artificial moral agents?

5. Inductive Logic programming, Bottom up Machine ethics
   Material: Chapter 19.5 of Russell und Norvig (2015), http://web.stanford.edu/~vinayc/logicprogramming/html/inductive_logic_programming.html
   Discussion: Tolmeijer u. a. (2020), Arkin (2008), Bentzen (2016), Malle u. a. (2015)
   Workshop: project idea proposals

6. Fairness, Accountability and Trust
   Material: Wieringa (2020), Bolukbasi u. a. (2016), Mehrabi u. a. (2019)
   Video: Arvind Narayanan tutorial at FAT2018
   https://www.youtube.com/watch?v=jIXIuYdnyyk
   Trusted AI and AI Fairness 360 Tutorial by Prasanna Sattigeri, September 18, 2019
   https://www.youtube.com/watch?v=IXbG2u4lOYI&feature=youtu.be
   Try out: https://aif360.mybluemix.net/

7. Explainability
   Material: Miller (2019a), Arya u. a. (2019), Gunning und Aha (2019)
   Video: Explainability 360 Tutorial by Amit Dhurandhar, September 18, 2019
   https://www.youtube.com/watch?v=TGPHPCg_zKA&feature=youtu.be
   Try out: https://aix360.mybluemix.net/

8. Responsible use of AI
   Material: Part III of Himma und Tavani (2008), Jobin u. a. (2019), Hagendorff (2019), Rahwan (2018)
   AI ISO standard in progress
   EU Ethics Guidelines for Trustworthy AI

The course was attended by 24 students and there were 5 student projects executed. The student evaluation format was retained with the same oral exam and project plan. In addition more intermediary assignments for the student projects were scheduled. These included: a refinement of research question, motivation, success criteria and methodology after the projects were decided, a related work analysis, and a two drafts deliveries. It is my judgement that the additional assignments of intermediary project reports has helped students with understanding scientific reporting, but it has also helped their argumentation skills. The grades for the projects were: A, A, B, C, C.

4.2 Observations

The order in which topics were thought could be improved in an AI Ethics course organisation. In INFO381-2020, machine ethics was introduced first and in hindsight this was not the right way. Machine ethics is the least intuitive of the AI Ethics sub-disciplines. Machine ethics requires understanding of moral philosophy, deontic logic and inductive logic programming. For most students this was a first introduction with moral philosophy with many new concepts that needed to be comprehended.

The machine ethics was followed with FAcct and XAI. Fairness and explainability require understanding of machine learning, but also accountability and transparency. Responsible use of AI was covered twice - in the firs class and in the last. As a topic, this one is most intuitive to the students, however, although it is not easy to see it does require a background of moral philosophy and also law and it is the most "further away" from computer and information science.

It is my conclusion that the modules should be thought effectively in reverse. First the students should be exposed to responsible AI, then the curriculum should follow with FAcct and XAI. This will expose the students to concepts from moral philosophy but in a programming environment that they are more familiar with. Once they have grasped values such as fairness, freedom, autonomy, and privacy they can be exposed to moral philosophy. The course will thus finish with machine ethics. Lastly, the

participation in discussions should be made a formal obligatory requirement to reinforce the value of argument building skills development.

5 AI Ethics as a bachelor course

The experience of INFO381 and the analysis of Fiesler u. a. (2020) will be taken into consideration in designing a bachelor course in AI Ethics. The Fiesler u. a. (2020) analysis identifies the challenge of integrating ethics into "purely technical programming courses", with many of the courses leaning towards high level and conceptual topics of impact of technology on society. I have also observed this as a challenge both in INFO381-2017 and INFO381-2020. Two of the tools that INFO381 relies on to immerse students into AI ethics cannot be transferred to a bachelor level course: self-selected group projects and studying from scientific articles.

Fiesler u. a. (2020) concluded that integrating ethical topics with programming "purely technical" courses "might even be a way to combat an 'I'm just an engineer' mindset that ethics is 'someone else's job'." The very open-ended group projects are the main tool that INFO381 uses to change this mind set and help future engineers and scientists see ethical AI impact as their own job. Bachelor students do not have the skill-set to execute and benefit from such projects.

One possible solution is to design a course that relies more heavily on topics of fairness and explainability with topics of moral philosophy interleaved. There is an added benefit to a FAccT and XAI "heavy" course. These areas concern machine learning especially. As the students will be following this course in parallel with a machine learning course, they can learn to consider ethical impact as part of machine learning, not as an afterthought. Practical programming exercises in assessing fairness, mitigating bias and generating explanations can be developed in parallel with the machine learning course.

While students should always be encouraged to read scientific articles, it would be inadequate to rely on them as the only study material for a bachelor course. We would need to be generate a script of selected texts using specifically systematic review articles and toolkit tutorials. A possible curriculum outline for bachelor AI Ethics, presuming 15 two hour units, is given in Table 2.

1.	Why AI ethics, why now
2.	Responsible AI and Norwegian AI guidelines
3.	What is an explanation?
4.	Accountability and Transparency
5.	Transparent vs black box machine learning
6.	Generating local explanations
7.	Argumentation and decisions
8.	What are good arguments and how to build them?
9.	Post-hoc explanations and surrogates
10.	How do we define bias and fairness?
11.	Data bias and mitigation
12.	Algorithmic fairness and mitigation
13.	Basics of moral philosophy
14.	Basics of moral philosophy
15.	An introduction to machine ethics

Table 2: Possible curriculum outline for a bachelor level AI Ethics course.

Lastly, it remains very important for the students to be thought how to analyse a problem that does not have an objectively correct solution. This again was done practically through the open-ended projects in INFO381. One possibility here is to execute the same self-selection of a project, but the students

would be tasked to write a popular science article communicating an issue in AI Ethics. They would be mentored in choosing a topic, discovering relevant material, and peer-evaluating the strength of the arguments in their own essays.

6 Conclusion

While there are global efforts to include courses on AI ethics in both bachelor and masters level, there is no clear consensus which topics such a course should focus on. An AI ethic course is difficult to fit in the beginning of a degree, because the students need to have a certain level of understanding of AI and computing methods to be able to comprehend their ethical impact. Students in computer and information sciences are rarely trained to handle a qualitative assessment of problems, in contrast to students in humanities and social science disciplines. Up until recently, computation and society were two separate environments and there was no need to interleave them.

INFO381-2017 and -2020 has had the goal to give an overview of the core issues and accomplishments in AI Ethics in a fashion that would motivate the students to pursue further learning in this area. Instead of the lecturer choosing one topic for in-depth study, the students were free to pursue their own interests and focus. It is critical to be aware that in young disciplines such as this we are educating the next generation of AI Ethics researchers and practitioners, and we need to motivate our students for this task, but also equip them with the basic skills they will need in their future careers.

The highly student-tailored approach of teaching AI Ethics in INFO381 has helped with most of the teaching AI Ethics challenges. However, the strategies used in INFO381 cannot be scaled to a bachelor level students who are not as equipped for independent learning as masters students. Teaching bachelor level AI ethics requires: 1) the preparation of a text book level material, 2) focus on the most "technical" and hands-on topics from AI Ethics, and 3) the gentle, strategic and practical introduction of argumentation and scientific communication skills.

References

[Algorithmic Watch 2020] ALGORITHMIC WATCH: *AI Ethics Guidelines Global Inventory*. https://inventory.algorithmwatch.org/. 2020. – [Online; accessed 20-May-2020]

[Amodei u. a. 2016] AMODEI, Dario ; OLAH, Chris ; STEINHARDT, Jacob ; CHRISTIANO, Paul F. ; SCHULMAN, John ; MANÉ, Dan: Concrete Problems in AI Safety. In: *CoRR* abs/1606.06565 (2016). – URL http://arxiv.org/abs/1606.06565

[Anderson und Anderson 2008] ANDERSON, Michael ; ANDERSON, Susan L.: ETHEL: Toward a Principled Ethical Eldercare System. In: *AI in Eldercare: New Solutions to Old Problems, Papers from the 2008 AAAI Fall Symposium, Arlington, Virginia, USA, November 7-9, 2008* Bd. FS-08-02, AAAI, 2008, S. 4–11. – URL http://www.aaai.org/Library/Symposia/Fall/2008/fs08-02-002.php

[Anderson 2008] ANDERSON, Susan L.: Asimov's "three laws of robotics" and machine metaethics. In: *AI Soc.* 22 (2008), Nr. 4, S. 477–493. – URL https://doi.org/10.1007/s00146-007-0094-5

[Arkin 2008] ARKIN, Ronald C.: Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture. In: *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*. New York, NY, USA : Association for Computing Machinery, 2008 (HRI '08), S. 121–128. – URL https://doi.org/10.1145/1349822.1349839. – ISBN 9781605580173

[Arya u. a. 2019] ARYA, Vijay ; BELLAMY, Rachel K. E. ; CHEN, Pin-Yu ; DHURANDHAR, Amit ; HIND, Michael ; HOFFMAN, Samuel C. ; HOUDE, Stephanie ; LIAO, Q. V. ; LUSS, Ronny ; MOJSILOVIĆ, Aleksandra ; MOURAD, Sami ; PEDEMONTE, Pablo ; RAGHAVENDRA, Ramya ; RICHARDS, John ; SATTIGERI, Prasanna ; SHANMUGAM, Karthikeyan ; SINGH, Moninder ; VARSHNEY, Kush R. ; WEI, Dennis ; ZHANG, Yunfeng: *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. https://arxiv.org/abs/1909.03012. 2019

[Awad u. a. 2018] AWAD, Edmond ; DSOUZA, Sohan ; KIM, Richard ; SCHULZ, Jonathan ; HENRICH, Joseph ; SHARIFF, Azim ; BONNEFON, Jean-François ; RAHWAN, Iyad: The Moral Machine experiment. In: *Nature* 563 (2018), Nr. 7729, S. 59–64. – URL https://doi.org/10.1038/s41586-018-0637-6. ISBN 1476-4687

[Bellman 1978] BELLMAN, Richard E.: *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd & Fraser Publishing Company, 1978

[Bentzen 2016] BENTZEN, Martin M.: The Principle of Double Effect Applied to Ethical Dilemmas of Social Robots. In: SEIBT, Johanna (Hrsg.) ; NØRSKOV, Marco (Hrsg.) ; ANDERSEN, Søren S. (Hrsg.): *What Social Robots Can and Should Do - Proceedings of Robophilosophy 2016 / TRANSOR 2016, Aarhus, Denmark, October 17-21, 2016* Bd. 290, IOS Press, 2016, S. 268–279. – URL https://doi.org/10.3233/978-1-61499-708-5-268

[Bolukbasi u. a. 2016] BOLUKBASI, Tolga ; CHANG, Kai-Wei ; ZOU, James ; SALIGRAMA, Venkatesh ; KALAI, Adam: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA : Curran Associates Inc., 2016 (NIPS'16), S. 4356–4364. – ISBN 9781510838819

[Bonnefon u. a. 2016] BONNEFON, Jean-François ; SHARIFF, Azim ; RAHWAN, Iyad: The social dilemma of autonomous vehicles. In: *Science* 352 (2016), Nr. 6293, S. 1573–1576. – URL https://science.sciencemag.org/content/352/6293/1573. – ISSN 0036-8075

[Bovens 2007] BOVENS, Mark: Analysing and Assessing Accountability: A Conceptual Framework1. In: *European Law Journal* 13 (2007), Nr. 4, S. 447–468. – URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0386.2007.00378.x

[Bransford u. a. 2000] BRANSFORD, John D. (Hrsg.) ; BROWN, Ann L. (Hrsg.) ; COCKING, Rodney R. (Hrsg.): *How people learn: Mind, brain, experience, and school*. Washington, DC : National Academy Press, 2000

[Caycedo Alvarez u. a. 2017] CAYCEDO ALVAREZ, M. ; BERGE, Ø. S. ; BERGET, A. S. ; BJØRKNES, E. S. ; JOHNSEN, D.V.K. ; MADSEN, F. O. ; SLAVKOVIK, M.: Implementing Asimov's First Law of Robotics. In: *Norsk Informatikkonferanse* (2017). – URL https://ojs.bibsys.no/index.php/NIK/article/view/396. – ISSN 1892-0721

[Chan 2019] CHAN, Dm: *Primary students to be taught AI in Guangzhou schools Government has decided to give priority to the development of AI, information and biopharmaceutical industries*. https://asiatimes.com/2019/07/primary-students-to-be-taught-ai-in-guangzhou-schools/. July 5 2019. – [Online; accessed 10-June-2020]

[Dennis u. a. 2016] DENNIS, Louise ; FISHER, Michael ; SLAVKOVIK, Marija ; WEBSTER, Matt: Formal verification of ethical choices in autonomous systems. In: *Robotics and Autonomous Systems* 77 (2016), S. 1 – 14. – URL http://www.sciencedirect.com/science/article/pii/S0921889015003000. – ISSN 0921-8890

[Dignum 2019] DIGNUM, Virginia: *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way*. Springer, 2019 (Artificial Intelligence: Foundations, Theory, and Algorithms). – URL https://doi.org/10.1007/978-3-030-30371-6. – ISBN 978-3-030-30370-9

[Fiesler 2018] FIESLER, Casey: *Tech Ethics Curricula: A Collection of Syllabi*. July 5 2018

[Fiesler u. a. 2020] FIESLER, Casey ; GARRETT, Natalie ; BEARD, Nathan: What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis. In: *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. New York, NY, USA : Association for Computing Machinery, 2020 (SIGCSE '20), S. 289–295. – URL https://doi.org/10.1145/3328778.3366825. – ISBN 9781450367936

[Foot 1967] FOOT, Philippa: The Problem of Abortion and the Doctrine of Double Effect. In: *Oxford Review* 5 (1967), S. 5–15

[Grosz u. a. 2019] GROSZ, Barbara J. ; GRANT, David G. ; VREDENBURGH, Kate ; BEHRENDS, Jeff ; HU, Lily ; SIMMONS, Alison ; WALDO, Jim: Embedded EthiCS: integrating ethics across CS education. In: *Commun. ACM* 62 (2019), Nr. 8, S. 54–61. – URL https://doi.org/10.1145/3330794

[Gunning und Aha 2019] GUNNING, David ; AHA, David: DARPA's Explainable Artificial Intelligence (XAI) Program. In: *AI Magazine* 40 (2019), Jun., Nr. 2, S. 44–58. – URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/2850

[Hagendorff 2019] HAGENDORFF, T.: The Ethics of AI Ethics - An Evaluation of Guidelines. In: *CoRR* abs/1903.03425 (2019). – URL http://arxiv.org/abs/1903.03425

[Himma und Tavani 2008] HIMMA, Kenneth E. (Hrsg.) ; TAVANI, Herman T. (Hrsg.): *The Handbook of Information and Computer Ethics*. John Wiley & Sons, Inc., 2008. – ISBN 9780471799597

[Jobin u. a. 2019] JOBIN, A. ; IENCA, M. ; VAYENA, E.: The global landscape of AI ethics guidelines. In: *Nature Machine Intelligence* (2019). – URL https://doi.org/10.1038/s42256-019-0088-2

[Johnson 1994] JOHNSON, Deborah: Who Should Teach Computer Ethics and Computers & Society? In: *SIGCAS Comput. Soc.* 24 (1994), Juni, Nr. 2, S. 6–13. – URL https://doi.org/10.1145/181900.181901. – ISSN 0095-2737

[Malle u. a. 2015] MALLE, Bertram F. ; SCHEUTZ, Matthias ; ARNOLD, Thomas ; VOIKLIS, John ; CUSIMANO, Corey: Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA : Association for Computing Machinery, 2015 (HRI '15), S. 117–124. – URL https://doi.org/10.1145/2696454.2696458. – ISBN 9781450328838

[McIntyre 2019] MCINTYRE, Alison: Doctrine of Double Effect. In: ZALTA, Edward N. (Hrsg.): *The Stanford Encyclopedia of Philosophy*. Spring 2019. Metaphysics Research Lab, Stanford University, 2019

[Mehrabi u. a. 2019] MEHRABI, Ninareh ; MORSTATTER, Fred ; SAXENA, Nripsuta ; LERMAN, Kristina ; GALSTYAN, Aram: *A Survey on Bias and Fairness in Machine Learning*. https://arxiv.org/abs/1908.09635. 2019

[Miller 2019a] MILLER, Tim: Explanation in artificial intelligence: Insights from the social sciences. In: *Artificial Intelligence* 267 (2019), S. 1 – 38. – URL http://www.sciencedirect.com/science/article/pii/S0004370218305988. – ISSN 0004-3702

[Miller 2019b] MILLER, Tim: "But Why?" Understanding Explainable Artificial Intelligence. In: *XRDS* 25 (2019), April, Nr. 3, S. 20–25. – URL https://doi.org/10.1145/3313107. – ISSN 1528-4972

[Ministry of Local Government and Modernisation 2020] MINISTRY OF LOCAL GOVERNMENT AND MODERNISATION: *The National Strategy for Artificial Intelligence*. https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=1. 2020. – [Online; accessed 20-May-2020]

[Moor 2006a] MOOR, James: The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. In: *AI Magazine* 27 (2006), Dec., Nr. 4, S. 87. – URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/1911

[Moor 2006b] MOOR, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: *IEEE Intelligent Systems* 21 (2006), Juli, Nr. 4, S. 18–21. – URL https://doi.org/10.1109/MIS.2006.80. – ISSN 1541-1672

[O'Neil 2017] O'NEIL, Cathy: *The Ivory Tower Can't Keep Ignoring Tech*. https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html. 2017. – [Online; accessed 20-May-2020]

[Pires Bjørgen u. a. 2018] PIRES BJØRGEN, E. ; ØVERVATN MADSEN, S. ; SKAAR BJØRKNES, T. ; VONHEIM HEIMSÆTER, F. ; HÅVIK, R. ; LINDERUD, M. ; LONGBERG, P.N. ; DENNIS, L.A. ; SLAVKOVIK, M.: Cake, Death, and Trolleys: Dilemmas as benchmarks of ethical decision-making. In: *Proceedings of the 2018 Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, AAAI/ACM, 2018, S. 23–29. – URL https://doi.org/10.1145/3278721.3278767

[Pistono und Yampolskiy 2016] PISTONO, Federico ; YAMPOLSKIY, Roman V.: Unethical Research: How to Create a Malevolent Artificial Intelligence. In: *CoRR* abs/1605.02817 (2016). – URL http://arxiv.org/abs/1605.02817

[Powers 2006] POWERS, Thomas M.: Prospects for a Kantian Machine. In: *IEEE Intelligent Systems* 21 (2006), Nr. 4, S. 46–51

[Rahwan 2018] RAHWAN, Iyad: Society-in-the-loop: programming the algorithmic social contract. In: *Ethics and Information Technology* 20 (2018), Mar, Nr. 1, S. 5–14. – URL https://doi.org/10.1007/s10676-017-9430-8. – ISSN 1572-8439

[Rawls 1958] RAWLS, John: Justice as Fairness. In: *The Philosophical Review* 67 (1958), Nr. 2, S. 164–194. – URL http://www.jstor.org/stable/2182612. – ISSN 00318108, 15581470

[Russell und Norvig 2015] RUSSELL, Steward ; NORVIG, Peter: *Artificial Intelligence: A Modern Approach*. 3. Pearson Education, 2015. – ISBN 0137903952

[Searle 1980] SEARLE, John R.: Minds, brains, and programs. In: *Behavioral and Brain Sciences* 3 (1980), Nr. 3, S. 417–424

[Sharkey und Sharkey 2012] SHARKEY, Amanda ; SHARKEY, Noel: Granny and the robots: ethical issues in robot care for the elderly. In: *Ethics and Information Technology* 14 (2012), Nr. 1, S. 27–40. – URL https://doi.org/10.1007/s10676-010-9234-6. ISBN 1572-8439

[Tolmeijer u. a. 2020] TOLMEIJER, Suzanne ; KNEER, Markus ; SARASUA, Cristina ; CHRISTEN, Markus ; BERNSTEIN, Abraham: Implementations in Machine Ethics: A Survey. In: *CoRR* abs/2001.07573 (2020). – URL https://arxiv.org/abs/2001.07573

[Turing 1950] TURING, Alan M.: Computing Machinery and Intelligence. In: *Mind* 59 (1950), Nr. October, S. 433–60

[Valland u. a. 2017] VALLAND, D. S. ; CARLSEN, T.A. ; LEA, M. ; PENSGAARD, A. ; KARLSEN, S. ; SÖTOV, P. ; SLAVKOVIK, M.: The Norwegian Oil Fund Investment Decider N.O.F.I.D. In: *Norsk konferanse for organisasjoners bruk at IT* 25 (2017), Nr. 1. – URL https://ojs.bibsys.no/index.php/Nokobit/article/view/409. – ISSN 1894-7719

[Vaughn 2014] VAUGHN, Lewis: *Beginning Ethics: An Introduction to Moral Philosophy*. W. W. Norton & Company, 2014

[Wallach und Allen 2008] WALLACH, Wendell ; ALLEN, Colin: *Moral Machines: Teaching Robots Right from Wrong*. USA : Oxford University Press, Inc., 2008. – ISBN 0195374045

[Wallach u. a. 2008] WALLACH, Wendell ; ALLEN, Colin ; SMIT, Iva: Machine morality: bottom-up and top-down approaches for modelling human moral faculties. In: *AI & SOCIETY* 22 (2008), Nr. 4, S. 565–582. – URL https://doi.org/10.1007/s00146-007-0099-0. ISBN 1435-5655

[Weizenbaum 1966] WEIZENBAUM, Joseph: ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. In: *Commun. ACM* 9 (1966), Januar, Nr. 1, S. 36–45. – URL https://doi.org/10.1145/365153.365168. – ISSN 0001-0782

[Weller 2017] WELLER, Adrian: Challenges for Transparency. In: *CoRR* abs/1708.01870 (2017). – URL http://arxiv.org/abs/1708.01870

[Wieringa 2020] WIERINGA, Maranke: What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, USA : Association for Computing Machinery, 2020 (FAT* '20), S. 1–18. – URL https://doi.org/10.1145/3351095.3372833. – ISBN 9781450369367

[Winfield u. a. 2014] WINFIELD, Alan F. T. ; BLUM, Christian ; LIU, Wenguo: Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In: MISTRY, Michael (Hrsg.) ; LEONARDIS, Aleš (Hrsg.) ; WITKOWSKI, Mark (Hrsg.) ; MELHUISH, Chris (Hrsg.): *Advances in Autonomous Robotics Systems*. Cham : Springer International Publishing, 2014, S. 85–96. – ISBN 978-3-319-10401-0