Teaching AI Ethics in a Flipped Classroom[*]

Greg Taylor[1], Author A[1], Debzani Deb[2], Author B[2]
[1]Department of Management, Marketing and MIS
[2]Department of Computer Science
Winston-Salem State University
Winston Salem, NC 27110
{taylorg,debd}@wssu.edu

Abstract

A flipped classroom approach can solidify AI ethics lessons in a few sessions. The approach described here introduces the Montreal Declaration of Responsible AI Development then asks students to apply it to a few case studies. Students post threads and responses to an on-line discussion board prior to a class session where student groups explore the cases in depth. Feedback and grades encourage high student engagement. Instructors could integrate similar AI ethics modules into any class where students have a minimal conceptual understanding of machine learning or AI. The learning objectives do not depend on the cases selected so new articles would be used over time to ensure student engagement. Instructors can easily modify the approach for use in an on-line setting.

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) based technologies play a crucial role in how we work, learn, communicate, and participate in society. As with many major scientific and technological breakthroughs, the use of AI and ML techniques has profound social and ethical implications. AI and ML technologies may reinforce racial and gender biases, perpetuate economic inequality, and violate privacy rights. These systems can target user beliefs and psychological traits with minimal transparency and accountability while eroding societal trust. Many academic, government, and industry efforts to develop principles for developing ethical AI have stressed the need for education among programmers, users and managers [1]. We use one broad-based effort, the Montreal Declaration for a Responsible Development of AI, as a framework for ethical analysis in the course modules described here [2].

Some universities have incorporated ethics contents into their AI/ML curricula specifically targeted toward the CS/CE majors [1]. While ethical training is essential for future developers of AI-enabled products, it is equally important for general practitioners and users of such systems. As educators, we need to weave such social and ethical considerations across all majors in all disciplines. We want to make sure these future practitioners can explore the ways AI/ML technology can have an impact on business stakeholders and their communities.

Instructors occasionally embed ethics in more technical curricula [3] [4] [5] [6], however, these courses are often not available to all university students. Another approach is to embed AI ethical issues into discipline-specific courses, ideally those targeting large groups of students or all students of a specific major. This ensures the ethical training is relevant but often requires training instructors in AI and ML. We used the latter approach. An instructor embedded an ethics module into a Business Analytics course required for all business-related majors. The approach could easily be adapted into many other disciplines that teach modeling or computing techniques.

Incorporating AI/ML related ethical training using this approach involves 1) teaching students some necessary ethical guidelines in a non-technical way and 2) providing them the opportunity to apply those guidelines to an AI-enabled situation, identify ethical issues, and assess potential trade-offs and solutions. In this paper, we present a novel modular approach for teaching societal and ethical implications of AI systems to non-majors with very limited prior technical and programming experiences. Ultimately, the goal is for students to recognize and describe ethical issues in real-world AI-based systems affecting their daily lives.

The AI ethics module presented in Section 2 introduces students to a principles-based framework developed in Dec 2018 by a group affiliated with the University of Montreal [2]. This framework describes 10 principles for responsible AI development. Students apply the ten principles to assess ethical problems with an AI/ML case-study in a flipped-classroom format [7] that facilitates collaborative learning. Students create and respond to threads in an on-line discussion forum before working in classroom-based groups to make short presentations on these topics. The instructor built two such case-studies by posting engaging questions in the forum and providing opportunities for learners to formulate answers (or sometimes to come up with inquisitive questions) independently and collectively.

Our empirical study is based on student performance and self-reflection survey data. They reveal that the on-line discussions act as catalysts to productive in-class conversations, persuasive arguments, and diverse viewpoints. These increase student engagement and academic performance. Our goal is to create a repository of case studies along with engaging questions and discussion activities.

The flipped-classroom approach could easily be adapted to an entirely virtual environment. Therefore, we anticipate that both the flipped-classroom modular approach and the repository of case studies will be useful in both traditional and virtual classrooms.

2 AI Ethics Module

We describe a modular, flipped-classroom approach to embedding AI Ethics instruction into a Business Analytics course in this section. The approach is well suited for courses focusing on modeling, computing or ethical issues in any discipline and is easily adapted to online learning. In the following subsections, we present learning outcomes, the flipped-classroom approach, and example case studies.

2.1 Learning Outcomes

We hope that students will be able to apply AI ethics frameworks to novel machine learning applications they encounter in their careers. Such frameworks will help them analyze, describe and suggest remedies for potential ethical violations. The specific learning outcomes for the presented module are

1. Learn about the Montreal Declaration for Responsible AI guidelines and apply the guidelines to recognize and describe ethical issues in AI-based systems. (LO1).

2. Discuss and reason, both alone and in collaboration with others, about the violation of the guidelines in an AI-enabled case-study and potential solutions of these violations. (LO2).

3. Gain enhanced awareness of approaches to minimize ethical problems that can arise in the development and implementation of AI-based systems. (LO3).

2.2 Flipped Classroom Approach

We followed a “flipped classroom” approach to teach the AI/ML ethics module. As an initial exercise, students read articles summarizing the Montreal Declaration for the Responsible Development of AI and efforts to build AI ethics boards. The module then utilizes two 75 minutes class sessions devoted to applying the framework to two AI-based applications (case study articles). Optionally, instructors can add more real-world case studies to explore other AI ethics issues with the Montreal framework.

Before each in-class ethics session, students contribute to an on-line discussion forum where they apply their knowledge of the ethical framework to assess a case study article describing an AI/ML application. Instructors assign students into one of two groups: each student in the first group creates a discussion thread and briefly describes and critically assesses the article based on engaging questions posed by the instructor. They also submit three questions or suggestions for in-class discussion. Each student in the second group then has one extra day to reply to one of the discussion threads created by a student in the first group. The instructor can reverse the roles of the two groups in subsequent case studies so each group has a chance to both assess the articles and respond to other student threads. At this phase, students worked independently.

The instructor then used the student-posed questions along with others to develop group discussion topics for an in-class session. Students split into groups of approximately five students. The instructor assigned each group a topic with a set of questions to orally present before the end of class. The groups had 15-20 minutes to formulate responses before presenting. Our classes were small enough to expect each student to contribute orally to the presentation and to respond to additional questions. Instructors could scale this for large sections by using parallel sessions supported by teaching assistants (TAs) and multiple presentation spaces. Instructors can encourage participation by grading each student’s discussion board response and oral contribution.

2.3 Case Studies

We utilized a case-study based approach to teach students about violations of ethical principles and possible solutions. In our modular intervention, we used two case studies based on popular newspaper articles. In this section, we detail our case studies as a template for use as is, or as inspiration for inclusion of other such case studies into the module. The first case study used a 2019 NY Times news article [8] detailing the facial recognition and tracking systems used by the Chinese government to monitor minority Uighur communities in Kashgar, China.

Half of the students described the monitoring system from the article, assessed it with the principles from the Montreal Declaration and posed three questions/topics for class discussion. Remaining students "added value" to their threads with answers to questions and/or expressing alternate viewpoints. Most students believed the Kashgar system violated all ten principles of the Montreal Declaration. A few questioned the neutrality of the article and wanted to read other viewpoints on this system. One important benefit of exposing students to these question/response activities is to allow them to think critically and independently about source information, ethical issues and possible solutions.

During the in-class discussion applying the Montreal Declaration principles to the Kashgar monitoring system, the instructor assigned a few principles for each group to apply to the AI-based monitoring system. Student participation was graded on critical analysis. Grades for individuals occasionally varied from the group.

A second case study given about four weeks after introducing the Montreal Declaration explored the ethical issues surrounding labeling and categorizing images. Students assessed Crawford and Paglen's claim that labeling and categorizing is inherently political [9]. The article cited ethically problematic examples from the ImageNet datasets [10]. Subsequently, half of the labels and categories in the dataset have been deleted.

The roles of the students were reversed from the first exercise above – half created threads assessing the article and listed three topics for discussion while the other half responded to those threads. The instructor did not directly prompt students to use the Montreal Declaration for analysis. Unfortunately, most students did not choose to use the Montreal Declaration as a framework for discussing the ethical issues posed by the new case study. With a bit of instructor prodding during the in-class group presentations, it dawned on a few students to use the Declaration to describe ethical issues. At this point, many saw the benefit of the using the principles to inform their analyses.

3 Experimental Evaluation

3.1 Results and Discussion

To evaluate the potential usefulness of the flipped classroom approach to learning AI ethics, we conducted a pilot study. We hypothesized students in the Business Analytics class could identify, describe and respond to ethical issues in AI/ML enabled systems. To test this hypothesis, we assessed the three learning outcomes (LO1, LO2, LO3) outlined in section 2.1 with student performance data. We also utilized an anonymous student experience survey (IRB approved) conducted at the end of the semester.

To evaluate LO1 (competency on the declaration principles and their applications), we used the student performance data on three ethics related multiple choice questions (MCQ) utilized in the final course exam. Twenty-five students enrolled but one student did not do any work. We eliminated that student from the statistics. Summarized responses from three exam questions are listed in Table 1. The results show the majority students (83%) were able to grasp the Montreal Declaration for AI guidelines (Q1.) More than half (54%) of the students were able to identify the violation of declaration principles in Amazon's automated firing of warehouse workers, a scenario not discussed in class (Q2.) Unfortunately, only 29% were able to correctly identify the three issues requiring removal of half of labels in the ImageNet dataset (Q3.) While 3 of the top 5 exam performers got the question correct, it was negatively worded (choose the non-issue) and the correct answer (replace human-generated labels with machine-generated) was not something discussed in class. The instructor should reword or replace Q3 in the future. Excluding Q3, about 68 percent of the students were able to correctly apply the ethical guidelines to recognize and describe ethical issues in AI-based systems (LO1). This result is consistent with observations during in-class labeling case study. Without prompting, students had difficulty applying the Montreal principles to a novel scenario. Perhaps additional case studies would improve future performance.

Table 1: MCQ Assessment in Final Exam

MCQ Question Learning Objective	Correct Responses (%)
Identify ethical principles from Montreal Declaration (Q1)	20/24 (83%)
Apply ethical principles to a new situation (Amazon's automated firing of warehouse workers.) (Q2)	13/24 (54%)
Identify issues requiring removal of half of labels in the ImageNet dataset. (Q3)	7/24 (29%)

We assessed the second learning objective, LO2 (ability to discuss ethical issues in person and in collaboration with others,) by utilizing two on-line discussion forum assignments and two 75-minute in-class group discussions devoted to discussing ethical issues related to the two case studies discussed earlier. Figure 1 shows the student performance data for the rubrics (A-F) described in section 2.3. "N" refers to students who did not submit or were not present. The surveillance system used in Kashgar, China case study is referenced as CS1. The image labeling and categorization procedures used for AI training sets case study is CS2. Most (76-84 percent) of the 24 students posted on the

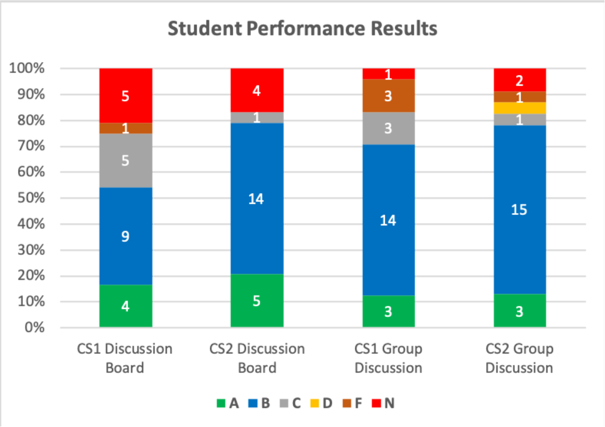

Figure 1: Discussion board and in-class group discussion results.

discussion boards prior to the corresponding in-class group discussion sessions. Higher percentages (84-92%) of students participated in group discussions and presented responses to questions during class.

Overall, 75-83 percent of the students earned C grades or better on the discussion board and in class on the two case studies. There was a small improvement in engagement on the second case study compared to the first. The grades on the two case studies made up 6 percent of the student's final grade in the course. Those who did not participate in either case study received a partial letter course grade less than they otherwise would have. The instructor may have improved engagement with this policy.

The grades of the engaged students also improved on the second study. Some may have been disappointed with "C" grades on the first case and recalibrated their expectations on the required work. Perhaps students learned to apply the ethical principles better after the first case. Overall, most students were engaged both in the on-line discussion forums and the in-class group presentations. Based on these assessments, most students met LO2.

We assessed LO3 (awareness of ethical issues in AI/ML systems) using two self-satisfaction end-of-course survey questions. Students responded using a 5-point Likert scale. Results are shown in Figure 2.

- Q1: I understand how the data science topics covered in this course could be utilized for societal good.
- Q2. I can discuss ethical issues surrounding the use of artificial intelligence in a professional setting.

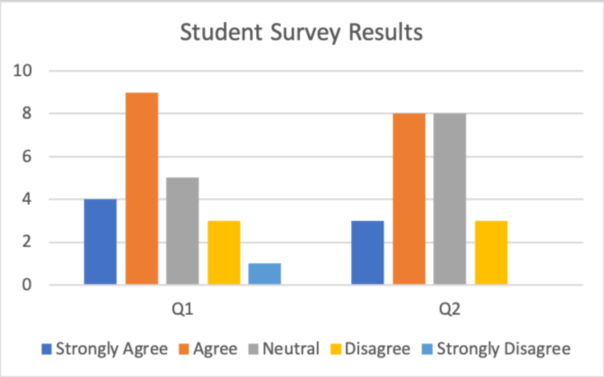

Figure 2: Student self-satisfaction survey results.

Of 22 questionnaire responses, only about half were confident about their ability to discuss AI ethical issues in a professional setting. A few more thought they understood how AI could be used for societal good. Fortunately, only a few disagreed with these statements. These questions produced more positive student feedback than 7 of the other 8 questions on the same survey. Most data science topics covered in the course involved statistical programming and the use of machine learning software. Based responses to the remaining 8 survey questions, we concluded that students are more confident about their knowledge of ethics than their ability to use software. Additional case studies might help more students achieve LO3.

3.2 Instructor Perspective

Based on these two sessions, the instructor thought the "A" students had a firm grip on the three learning objectives. They would be able to utilize guidelines in a professional setting to minimize ethical problems that could arise in the development and implementation of proposed AI systems. All students realized AI systems can have ethical problems.

Applying the Montreal Declaration to labeling and categorization without prompting was difficult for students after a month had passed since introducing the topic. Still, the goal is for students to recognize ethical problems with AI systems and use the Declaration to help describe these issues. Achieving this objective would have required at least one more ethical case study for these students.

4 Conclusions and Discussion

The flipped classroom used in this study requires students to participate in an on-line discussion board before attending an in-class session. It is an effective way to learn AI ethics. This approach allows students to assess AI-based systems with a set of ethical principles in as few as two in-class sessions. Once an ethical framework such as the Montreal Declaration for Responsible AI Development is introduced, applications from the popular media or academic literature may be assessed.

New applications appear frequently in these outlets and instructors can substitute them for the Uighur monitoring and labeling cases described above. One could easily construct cases to achieve the same learning objectives from descriptions of self-driving cars, social media monitoring, and other topics frequently explored in the popular press. Changing the cases from year to year limits opportunities for students to inappropriately use the work of others. Students must engage and add value to the discussion boards and in-class presentations to achieve the learning outcomes and receive positive instructor feedback.

The flipped classroom approach can facilitate discussions of AI ethics into any course. Students needed minimal prior knowledge to analyze the Kashgar case study. However, students need a conceptual understanding of machine learning before exploring some ethical issues. Appreciating the ethical issues involved with ImageNet requires some background in training machines, finding examples, labeling, modeling and predicting. The ImageNet labeling case can reinforce student's conceptual understanding of machine learning, especially in non-technical courses. Instructors could easily develop similar cases to reinforce other AI concepts.

The flipped classroom can be easily adapted to on-line learning environments. On-line instructors could replicate the in-class group discussions and presentations with the help of a synchronous conferencing tool. Chat facilities with breakout rooms can substitute for in-class group work. A moderator could replace in-class group presentations with chat responses or video presentations. In an asynchronous on-line environment, student groups could produce a written response to assigned topics on a discussion forum or similar venue.

5 Acknowledgements

This research was supported by National Science Foundation Award 1600864. The grant provides encouragement and training to embed data science topics and courses into curricula.

References

[1] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. Artificial intelligence for social good: A survey. *arXiv*, 2020.

[2] Montreal declaration for a responsible development of artificial intelligence, 2018. https://www.montrealdeclaration-responsibleai.com/reports-of-montreal-declaration.

[3] E Burton, J Goldsmith, S Koenig, B Kuipers, N Mattei, and T Walsh. Ethical considerations in artificial intelligence courses. *AI Magazine*, 38(2):22–34, 2017.

[4] Tom Williams and Qin Zhu. An experimental ethics approach to robot ethics education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. http://par.nsf.gov/biblio/10125972.

[5] Heidi Furey and Fred Martin. AI education matters: A modular approach to AI ethics education. *AI Matters*, (4):13–15, 2019. https://doi.org/10.1145/3299758.3299764.

[6] Barbara J. Grosz, David Gray Grant, Kate Vredenburgh, Jeff Behrends, Lily Hu, Alison Simmons, and Jim Waldo. Embedded EthiCS: integrating ethics across CS education. *Communications of the ACM*, 62(8):54–61.

[7] Michael L. Wallace, Joshua D. Walker, Anne M. Braseby, and Michael S. Sweet. "now, what happens during class?" using team-based learning to optimize the role of expertise within the flipped classroom. *Journal on Excellence in College Teaching*, 25(3):253–273, 2014. http://152.12.30.4:2048/login?url=https://search.proquest.com/docview/1651854490?accountid=15070.

[8] Chris Buckley, Paul Mozur, and Austin Ramzy. How china turned a city into a prison. *The New York Times*. https://www.nytimes.com/interactive/2019/04/04/world/asia/xinjiang-china-surveillance-prison.html.

[9] Kate Crawford and Trevor Paglen. Excavating AI: The politics of images in machine learning training sets, 2019. https://www.excavating.ai/.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. 2009. http://www.image-net.org.