10417/617 Intermediate Deep Learning: Ethics Module

Course Website: https://andrejristeski.github.io/10417-20/
Instructor: Professor Ruslan Salakhutdinov
Ethics Module Student Developers: Nadine Bao and Sreedhar Radhakrishnan
Approximate Class Size: 60 - 70 students
Total Workload: 3.5 - 5 hours

Learning Objectives:

At the end of the module, students should be able to:

- Recognize ethical considerations and bias introduced in sequence-to-sequence models, particularly language models.

- Analyze ethical and psychological aspects of deep reinforcement learning with a focus on machine ethics.

- Identify ethical consequences of students' deep learning applications as an add-on component to the course project.

Content Justification:

The primary purpose of our module is to introduce students of 10-417/10-617 to topics in ethics related to the technical subjects they learn throughout the semester, mainly language models and deep reinforcement learning.

The module follows a plug-and-play design and can be seamlessly integrated into the students' existing coursework without drastically increasing their workload. To allow for thoughtful consideration of the topics in ethics presented, students will have the opportunity to engage in two discussion formats: class-wide and small groups. We find that in-class student and professor-led discussions are a good change of pace for courses that are generally very technical, as it allows for interaction among students and improves engagement. Students will be required to complete reading and journal assignments before class to help them contribute meaningfully during in-class discussions.

Additionally, towards the end of the semester, students will be asked to submit a reflection on their final projects which prompts them to consider the ethical implications of their creations. Considering the existing workload students may already have under this and other courses, we have designed all take-home assignments to take under an hour to complete.

Acknowledgements

We would like to thank Professor Illah Nourbakhsh and Victoria Dean for giving us the opportunity to contribute to machine ethics pedagogy. We would also like to thank Professor Ruslan Salakhutdinov for supporting us during the course of the module development and providing us with the creative freedom to share and implement our ideas. This was a great learning experience for the both of us.

Sincerely,
Nadine and Sreedhar

Table of Contents

Note:
Suggested rubrics are located after each assignment that needs to be submitted.

Class Outline and Instructions for Instructors

Readings and In-Class Activity:

Language Models Ethics Reading and Journal

(40 minutes, due prior to class | Possible Assignment Dates Range: September 15 - September 30)

This assignment should be assigned to students prior to **Ethics: Language Models Lecture**. Students should complete and submit this assignment prior to class in order to gain some background on the topics that will be discussed during the lecture. The article the students will read is “We read the paper that forced Timnit Gebru out of Google. Here’s what it says.” by Karen Hao. It belongs to the MIT Technology Review which does have a paywall; however, MIT Technology Review does allow a limited number of free stories prior to requiring a subscription. We have also included a pdf copy of the article in the case that the article can no longer be accessed.

Ethics: Language Models Lecture

(30 minute lecture time + 15 minute discussion | Possible Assignment Dates Range: September 15 - September 30)

The lecture slides include a brief overview of **Language Models Ethics Reading and Journal** assigned to students prior to class and further explores ethical considerations in language models.

Prior to the lecture, the instructor should complete the following:

- Read the article: “We read the paper that forced Timnit Gebru out of Google. Here’s what it says.” by Karen Hao.
- Make any desired changes to slides to reflect any additional content the instructor would like to include.
- Take note of anecdotes or analogies the instructor would like to share with students during the lecture.
- We have provided guidance for each slide in **Supporting Slides Instructor Aid** with further explanation of points mentioned in the presentation.

Deep Reinforcement Learning Reading/Journal and In-Class Activity

(40 minute reading/journal prior to class + 20 minute in-class discussion + 20 minute class-wide discussion | Possible Assignment Dates Range: November 20 - November 27)

This assignment should be assigned to students before the **In-Class Activity** for Deep Reinforcement Learning. Students should complete and submit this assignment prior to class in order to gain some background on the topics that will be discussed during the

In-Class Activity. The article the students will read is “Machines That Don’t Kill: How Reinforcement Learning Can Solve Moral Uncertainties.”

The **In-Class Activity** involves dividing the class into groups of 5 (use breakout rooms if the class is online) and providing the students with 20 minutes to engage in peer discussion and healthy debate to discuss their viewpoints. The instructor can choose to invite two student groups at random to present their views (10 minutes each) which includes time for questions and discussion from other group members and peers.

Class Project:

Assignment Due Date: Same as the due date of the class project (to be decided by course staff)

Project Reflection: Concept Diagram (20 minutes):

This assignment is intended to be integrated into the final project. It can be completed with the respective project team members. The assignment should be given out at the beginning of the final project to prepare students for the workload but is intended to be completed near the end of the final project as a way to reflect on the potential ethical implications of the students’ work. Students may opt to submit hand drawn diagrams or create their diagrams digitally using tools such as Google Drawings. The concept diagram should be included in the final project report under ‘Project Reflection’.

Project Reflection: Assignment Question (30 minutes):

This assignment is intended to be integrated into the final project. It should be completed individually. The assignment should be given out at the beginning of the final project to prepare students for the workload but is intended to be completed near the end of the final project as a way to reflect on the potential ethical implications of the students’ work. The assignment question should be included in the final project report under ‘Project Reflection’.

Project Reflection Extra Credit: Programming Question (90 minutes)

This assignment is intended to be integrated into the final project. It can be completed with the respective project team members. The assignment should be given out at the beginning of the final project to prepare students for the workload but is intended to be completed near the end of the final project as a way for students to learn about testing their application for bias. The extra credit question (if completed) must be included in the final project report under ‘Project Reflection’.

Language Models Ethics Reading and Journal

Please read “We read the paper that forced Timnit Gebru out of Google. Here’s what it says.” by Karen Hao and submit your journal responses to the discussion questions below prior to class. Your journal response should be at least 300 words and reflect thoughtful consideration of the questions given.

Optionally, you can also read the paper in question: “On the Dangers of Stochastic Parrots:Can Language Models Be Too Big?”

Link to article:
https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/
Link to paper in question:
https://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

Discussion Questions:

1. Should applications of language models be required to adhere to current moral customs? Who, if anyone, should be responsible for meeting this requirement?

2. Who ultimately suffers from the negative environmental impacts from training language models?

3. Are you concerned about the privacy of your data and information given that it could potentially be used to improve language models?

Rubric
Assignment: Language Models Ethics Reading and Journal
Total possible points: 6 pts

Criteria	Ratings				
Depth of Analysis	**6** - Student answers all discussion questions - Responses show thoughtful consideration.	**5** - Student answers all questions, but arguments lack thoughtful consideration and depth.	**4** - Student answers all questions but fails to address all parts of the questions.	**3** - Student only answers a subset of the questions. **or** - Student submits a response that falls under 150 words.	**0** - No submission **Or** - Very few words

Deep Reinforcement Learning Reading/Journal and In-Class Activity

Deep Reinforcement Learning Reading/Journal

Please read “Machines That Don’t Kill: How Reinforcement Learning Can Solve Moral Uncertainties” as a preparation for the in-class activity involving peer discussion of the paper. Please submit your response to the discussion questions prior to class. Your response should be at least 300 words and reflect thoughtful consideration of the questions given below:

Link to the reading material:
https://analyticsindiamag.com/reinforcemenet-learning-moral-dilemma-ethics/

Discussion Questions:

1. How would you train a reinforcement learning agent to handle a situation such as the trolley problem described in the article? How should a self-driving car handle such a situation?
2. What are some ways the gap between moral philosophy and machine ethics can be bridged?

Rubric
Assignment: Deep Reinforcement Learning Reading/Journal
Total possible points: 6 pts

Criteria	**Ratings**				
Depth of Analysis	**6** - Student answers both discussion questions - Responses show thoughtful consideration	**5** - Student answers all parts of both questions, but arguments lack thoughtful consideration and depth.	**4** - Student answers both questions but fails to address all parts of the questions.	**3** - Student only answers one of two questions assigned. **or** - Student submits a response that falls under 150 words.	**0** - No submission **Or** - Very few words

In-Class Activity

The reading assignment submission by the student ensures that the student comes prepared for the class discussion. Divide the class into groups of 5 (use breakout rooms if the class is online). The discussion questions are intentionally open ended. Provide the students with 20 minutes to engage in a peer discussion and healthy debate to discuss their viewpoints. Invite 2 groups at random to present their views (10 minutes each) which includes time for questions and discussion from other group members and peers.

Supporting Slides Instructor Aid

We have prepared slides to aid with the Language Models Ethics Reading class discussion. The instructions/aid for using the slides have been provided below:

Slide 1: Ethics: Language Models

This slide can serve as a quick introduction to let students know that this lecture is related to the **Language Models Ethics Reading and Journal** they submitted prior to class.

- Example: Today in class, we will be expanding on the reading and discussion questions we assigned to you. At the end of the lecture, there will be an opportunity for you to share some thoughts on the topics discussed today with the rest of the class.

Slide 2: Reading Overview

This slide serves as a refresher for students to recall some basic ideas that were addressed in the article.

In summary, a paper that Google AI Ethicist Timnit Gebru contributed to was not approved for publication. The reason for the rejection provided by the head of Google AI was that the paper failed to reference sufficient works related to resolving bias and improving energy efficiency in large language models. The article mentions four main dangers of large language models that are addressed in the paper:

1. The carbon footprint of training models increases at an alarming rate as the amount of data in models increases.
2. The incentive to collect more and more data from the Internet has led to the inclusion of problematic language in training data.
3. Researchers have pointed out that more efforts need to be put into models that aim to better understand language rather than models that aim to better manipulate language (the latter is more profitable).
4. Language models that are able to produce convincing language are abused in order to quickly spread misinformation. Additionally, mistranslations also lead to the spread of misinformation.

Slide 3: Academic Freedom and Agency

Many critics of the situation view Google's disapproval of the paper and Timnit Gebru's departure from the company as a form of research censorship. Critics convey that Google did not wish to have the paper published because it exposed the negative impacts of the lucrative language models it continues to use and develop. This leads us to a topic in ethics: agency. Agency is a person's ability to make their own decisions and act independently.

In the situation described in the article, the researchers met an obstacle when attempting to publish their work. The academic freedom and agency of the researchers can be seen as limited, as they cannot openly publish without approval from a higher authority.

So should researchers work at companies where their academic freedom may be limited? This is a difficult question to answer because researchers tend to want to work for large companies where they will receive higher compensation.

Slide 4: Privacy Concerns

This slide segways into an ethical consideration in language models not mentioned in the reading. It is designed to be an interactive slide during which the professor will seek responses to the questions from students attending lecture. To prevent discussion from taking too much of the lecture time allocated time, we recommend about two student responses per question. Alternatively, the instructor can ask if any students would like to respond to the questions of their own choosing.

The questions are listed below and on the slides:

1. What data is used to train language models, and where is it stored?
2. Are your conversations with virtual assistants being saved?
3. What are some of the privacy concerns in the language modelling cycle?
4. Can we build language models without raising privacy concerns? How?

Slide 5: Environmental Impact

This slide dives into the environmental impact of training language models. Like slide 4, it is designed to be an interactive slide during which the professor will seek responses to the questions from students attending lectures. To prevent the discussion from taking too much of the allocated time, we recommend about two student responses per question. Alternatively, the instructor can ask if any students would like to respond to questions of their own choosing.

The questions are listed below and on the slides:

1. Are the accuracy gains of training large language models worth the carbon footprint produced?
2. Many energy sources that drive cloud infrastructure and other systems that power language model training are not carbon neutral. What are some ways to transition towards renewable sources of energy?
3. What are your thoughts on the environmental impact of language models? Is it an immediate concern?

Slide 6: More Discussion Time!

To conclude, slide 6 discusses certain questions asked in the homework as well as other questions to engage the student to think about the ethical considerations of language models. We recommend the instructor provides his/her point of view of the discussion questions and then asks 2-3 students for their thoughts and responses. The questions are listed below and on the slides:

1. Should applications of language models be required to adhere to current moral customs? Who, if anyone, should be responsible for meeting this requirement?
2. Who ultimately suffers from the negative environmental impacts from training language models?
3. Are you concerned about the privacy of your data and information given that it could potentially be used to make language models better?

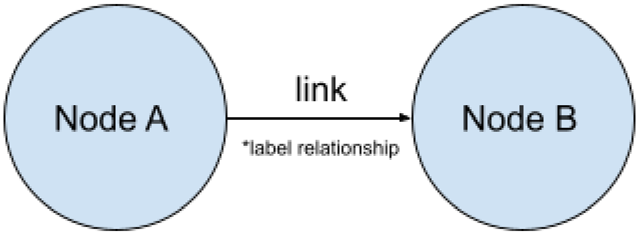
Project Reflection

As a part of the course project students are required to build a deep learning application to apply their learnings. The **Project Reflection** involves an extension of the course project wherein students are required to reflect on the ethical considerations of their work. This includes a concept diagram as well as a written section in their final project report. There is also an optional extra credit programming assignment involving writing test cases to identify bias in their project.

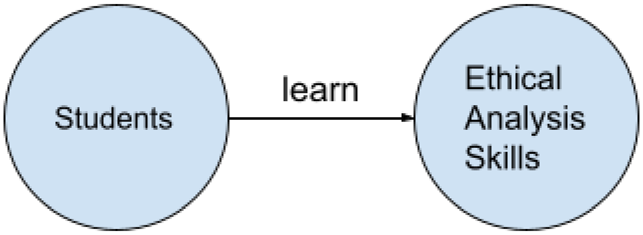
Project Reflection: Concept Diagram

Construct a concept diagram that illustrates the relationship between your project and other entities that are in some way impacted by your project. The concept diagram should be submitted under 'Project Reflection' in your final report.

A concept diagram is constructed using **nodes** and **links**.



Nodes represent groups, individuals, issues, and things that are impacted by your project. **Links** are arrows that express the directional relationship between different nodes. For example, to express the relationship between students(Node A) and ethical analysis skills(Node B), we can draw:



Notes:

- Sometimes links/relationships can be bidirectional. We can illustrate such scenarios using double-headed arrow (↔).
- Additionally, not all nodes will be directly linked together as they may not directly impact each other.

Requirement:

- Your concept diagram must have at least **7 nodes**, including your project as 1 node.
- **Draw links** between nodes that share a relationship.
- **Label links** with the relationship they represent.

Rubric
Assignment: Project Reflection: Concept Diagram
Total possible points: 6

Criteria	**Ratings**		
Nodes	**2** - Student has all 7 required nodes. - Nodes chosen show thoughtful consideration from the student.	**1** - Student has less than 7 nodes but more than 3 nodes. **Or** - Student has 7 nodes, but the nodes chosen lack depth.	**0** - Student has 3 or less nodes.
Links	**2** - Links are thoughtfully selected and clearly convey relationships between nodes.	**1** - There are very few links despite potential relationships between student’s nodes.	**0** - Student is missing links.
Relationships	**2** - Student expresses a variety of relationships.		**0** - Student did not label relationships.

Project Reflection: Assignment Question

Provide your response to the question stated below. Your response should be at least 300 words and reflect thoughtful consideration of the question given below:

“Did you think about the ethical considerations of your deep learning application while designing your system? What are some of the moral issues/concerns of your application when deployed at scale?”

Rubric
Assignment: Project Reflection: Assignment Question
Total possible points: 6 pts

Criteria	**Ratings**				
Depth of analysis	**6** - Student answers all discussion questions - Responses show thoughtful consideration	**5** - Student answers all questions, but arguments lack thoughtful consideration and depth.	**4** - Student answers all questions but fails to address all parts of the questions.	**3** - Student only answers a subset of the questions. **or** - Student submits a response that falls under 150 words.	**0** - No submission **Or** - Very few words

Project Reflection Extra Credit: Programming Question

You have built a deep learning system trained on a large dataset. Can you write unit tests to ensure that the results are as expected? Can you identify a bias in your model? If you did not identify a bias, but tried to find one, highlight the steps you took to search for the bias and write a note on why you feel your model does not show bias. This could include explanations involving the data variety and source, training methodology, and use case.

Rubric
Assignment: Project Reflection Extra Credit: Programming Question
Total possible points: 10 pts

Criteria	**Ratings**			
Effort	**10** - Student writes unit tests to ensure robustness of the application. - Student writes test and identifies bias in the system **or** student highlights steps taken to identify bias and justifies lack of it.	**8** - Student writes unit tests but does not identify bias in the system and fails to justify the result.	**5** - Student theoretically justifies presence or absence of bias in the system but does not write test cases to identify it.	**0** - No submission