Ethics and Robotics - Spring 2021 Final Project

Ethics Incorporation

10-315: Introduction to Machine Learning

Dani Grodsky & Adya Danaditya

Table of Contents

Executive summary

This document is a manual for the ethics incorporation module that we prepared for the 10-315: Introduction to Machine Learning class. The module's goal is for the students to be able to articulate ethical considerations in machine learning applications, along with other accompanying ethics learning objectives tied with concepts existing in the class syllabus. To achieve this goal we designed a module that would run throughout the semester, with a short introduction and a trigger topic in the first lecture that would serve as a starting point, followed by ethics learning activities like discussion board reflections and podcast listening embedded to some class topics and culminating in a practical activity in which students peer review Machine Learning innovations they created themselves in an ethics perspective. This design is derived from constraints that are stated by the cooperating instructor and principles we tried to apply based on our personal observations.

Background

Class Overview

10-315: Introduction to Machine Learning, which will be taught by Professor Pat Virtue in Fall 2021, is a fully undergraduate course with around 100 students usually enrolled. There are two 80 minute lectures and one 80 minute recitation each week. The students are usually second or third years who have taken other courses in probability and mathematical foundations of computer science as prerequisites. In previous iterations of this and other classes in similar veins, Professor Virtue has used active learning techniques like think-pair-share, in-class polls and breakout room discussions in-person or during remote instruction.

Relevant Topics & Themes

The class consists of introductions to various machine learning concepts and techniques, from decision trees to neural networks, elaborated in Table 1 below.

Week 1	Introduction to Classification, Regression & ML Concepts
Week 2	Linear Regression
Week 3	Probabilistic Linear Regression
Week 3	Logistic Regression
Week 4	Regularization
Week 5	Naive Bayes
Week 5	Generative Models
Week 6	Neural Networks
Week 7	Nearest Neighbor
Week 8	Decision Trees
Week 9	Cross-validation & Nonparametric Regression
Week 10	SVM
Week 11	Dimensionality Reduction
Week 12	Recommender Systems
Week 13	Clustering

Week 14	Learning Theory
Week 15	Ensemble Methods

Table 1: Weekly topics taken from the syllabus

The team and the instructor have looked through each of the planned topics in the syllabus to find relevant pairings between ethical concepts and the corresponding machine learning techniques. Through this process, we found several entry points to ethics that we can develop, including explainability in ML models, the influence of recommender systems, and consequences of dataset bias (including representativeness, fair data collection, etc).

Design

In designing the ethics integrations there are some things we kept in mind constraints - including logistical or personal preferences of the instructor - as well as some key principles that the we felt could add value to the student's ethics learning experience.

Constraints

Logistical issues regarding class

- Given the large class size, student-led presentations or discussions are more time consuming and difficult
- There is a lot of material in every class so any material added needs to be thought through and appropriately condensed
- Uncertainty about whether the class will be in person, remote, or a hybrid

Instructor-related constraints

Professor Virtue has prior experience in adding ethics modules to his classes and has tried some models - including having a standalone class at the end, which given the time in the semester did not seem to be as effective as possible. He also noted some constraints apart from technical and logistical challenges, which include the desire to have specific cases and questions to help guide any in-class discussion - to reduce any need to spitball - as well as the desire to find strategies for adding a quantitative grading component to the ethics materials.

Principles

Principles the team tried to infuse into the learning process

- Incorporate active learning techniques
  This is done to increase student involvement and enable students to engage in higher-level activities in the Bloom Taxonomy, in some way ensuring greater levels of comprehension
- Ethics infused as a relevant add-on to the already existing materials, not a separate, disjunctive topic
  Preventing confusions that may arise due to the heavily contrasting tone of ethics teaching and CS teaching
- Grounded in real-world cases
  Adding stakes and urgency in the discussions rather than limiting it to abstract concepts
- Highlighting ethics as a necessity instead of a voluntary topic
  Promoting the culture of a hands-on and prioritized approach in considering and taking direct action related to technological consequences - both positive and negative

Design Strategies

Here are some strategies that we use to design the implementation of these ethics integration - a byproduct of the principles and the constraints that are set in this project and mentioned earlier

Strategies	**Reasoning**
Will provide specific research paper-based cases and questions so that there are bounds and guidelines around the conversation	Prevent spitballing The use of examples and case studies provides more detail and grounding
Will use a portion of the 5% participatory grade structure already present in the class and a portion of standalone homework scores (if there's an ethics question tied in that particular homework). The activities will be graded based on completion and a simple check of relevance.	Instructor's uncertainty around measurement

Have smaller ethics inclusions throughout the semester (in-class discussion and/or homework)	Easier scheduling Builds the habit of connecting critical ethics consideration to ML development
Leveraging digital tools - such as discussion boards - for activities	Make up for the lack of in-person discussion/meetings and presentation that would be harder given the large class size

Table 2: Design strategies

Implementation

Overview

The ethics integration module we design would consist of these parts:

1. Short chunk in the intro lecture that:
    a. Touch on logistics & teases upcoming activities regarding ethics
    b. Touches the general topic of dataset bias as a trigger topic
2. Learning activities (homework and/or class) on ethical tie-ins relevant to two chosen class topics (Explainability in Decision Trees and Recommender Systems)
3. A culminating activity in the last class to wrap up and solidify the learnings, which consists of an ML-related Innovation Pitch + Ethics Peer Review

Detail Per Section

Introduction Session

Plan: This session will introduce the ethics theme and visual that will pop up throughout the semester. It will also include a brief discussion of data collection and how, for example, representation (or a lack thereof) in the initial dataset can influence the algorithmic results

Learning Objective: Assess how characteristics of the dataset and its collection can affect analysis outcomes

Activity: This introduction session will take some part of the class' first lecture, with some open-ended questions for members of the class to raise their hand to answer if willing

Supporting Materials

- See ***Appendix 1*** for slides with notes (the Powerpoint file is also shared in Google Drive)

- Rogati, M. (2017, June 12). *The AI Hierarchy of Needs*. Hackernoon. https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007
- Goodman, R. (2018, October 12). *Why Amazon's Automated Hiring Tool Discriminated Against Women*. ACLU. https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against
- Lashbrook, A. (2018, August 16). *AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind*. The Atlantic. https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/
- Haenssle, H.A., et al. (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol.* 29(8):1836-1842. doi:10.1093/annonc/mdy166.

Possible Extensions

There are other types of dataset bias that could be discussed, included in the first homework or in a class/homework later in the semester. Some possible resources for this additional discussion include:

- Anonymous Authors (2020) Dataset Bias in Diagnostic AI Systems: Guidelines for Dataset Collection and Usage. https://doi.org/10.1145/1122445.1122456
- Williams, B.A., et al. (2018) How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions and Policy Implications. Penn State University Press. https://www.jstor.org/stable/10.5325/jinfopoli.8.2018.0078

Ethics Learning Activity: Explainability

Plan: Introduce the theme of explainability and its complexity within the lecture, then have the students reflect more deeply and apply the concept of explainability through further reading and a discussion exercise

Learning Objective: Define the concept of explainability as well as its consequences in a real-world application

Activity:

1. Introduce the theme of explainability via lecture (lecture slide included)
2. Briefly discuss trade-offs and areas of contention in the domain of explainability
3. Have the students read accompanying material on explainability and reflect on the benefits as well as the consequences when it's lacking in the context of health, financial inclusion, criminal justice, and science via a discussion board post (likely through Piazza but may change in the future)

Supporting Materials

- See ***Appendix 2*** for slides with notes (the Powerpoint file is also shared in Google Drive), included within are the prompts and detailed instructions of the discussion board task
- The Royal Society (2019). Explainable AI: The Basics - Policy Briefing [White paper]. The Royal Society.
- Referenced materials in slide:
    - Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. MDPI AG. Retrieved from http://dx.doi.org/10.3390/e23010018
    - Sendak, M., Elish, M., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., & O'Brien, C. (2019). "The human body is a black box": Supporting clinical decision-making with deep learning. ArXiv:1911.08089 [Cs]. http://arxiv.org/abs/1911.08089
    - Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
    - MI in Healthcare Workshop Working Group, Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., & Mandl, K. D. (2020). Machine intelligence in healthcare—Perspectives on trustworthiness, explainability, usability, and transparency. *Npj Digital Medicine*, *3*(1), 47. https://doi.org/10.1038/s41746-020-0254-2

Ethics Learning Activity: Recommender Systems

Plan: Within the lecture, briefly introduce the lens of ethics to the topic of recommender systems and then include a homework activity to help students continue to think more deeply about these ethical considerations

Learning Objective: Explain ethical considerations related to the creation and use of recommender systems

Activity:

1. Lecture component with one slide introducing the topic and one slide introducing the homework activity;
2. Have students listen to an episode of the Toward Data Science podcast on *Ethical Problems with Recommender Systems* (January 2021) and answer provided prompts in the online discussion board

Supporting Materials

- See ***Appendix 3*** for slides with notes (the Powerpoint file is also shared in Google Drive)
- See ***Appendix 4*** for homework write-up with discussion board questions (word document also shared in Google Drive)
- Harris, J. (2021, January 27) Ethical Challenges of Recommender Systems. Toward Data Science. https://towardsdatascience.com/ethical-problems-with-recommender-systeems-398198b5a4d2
- Milano, S., et al. (2020) Recommender systems and their ethical challenges. AI & Society. 35, 957-967. https://doi.org/10.1007/s00146-020-00950-y
    - Articles cited in this paper and those that cite this paper are great resources for possible extension

---

Final Class: 'ML for Good Activity:

Plan: Run a group activity with students aimed at inventing Machine Learning applications for social good that includes breakout rooms, the discussion board, peer review, and team presentations. The full activity should take up about half of each of the last two classes, leaving time to wrap up other class logistics and/or expand the ethics discussion

Learning Objective: Generate applications of Machine Learning for social good and summarize ethical considerations of your own and ideas and that of others

Activity:

Will be split into two classes, the penultimate class and the last class

Penultimate class activity

- Introduce the activity in the lecture/slides
    a. Students will be given 20 minutes to work in teams, brainstorm machine learning applications for social good, and ultimately choose one to write a short pitch about (see part d)
    b. Students will be randomly assigned to (or potentially allowed to choose) teams and/or breakout rooms (depending on if the class is in-person or virtual)*
    c. In a shared Google Sheet, each team will sign up to focus on one of six social good areas - healthcare, privacy, financial health, sustainability, education, or civic engagement (built into the Google Sheet is a max number of teams per category)
    d. The pitch each team creates should be a 6-8 sentence summary of the idea, including a brief plan for data collection and model implementation; one member of each team will post the pitch on the Piazza discussion board during or after class. The idea does not need to be something the students can build themselves (yet).
    e. After class (based on the logistics of the topic sign-up**), each group will be paired with another group for whom they will read their pitch on Piazza and begin to think about additional ethics considerations of the other team's idea

*Forming teams: Ideally there will be an even number of teams signed up for each social good category (the Google Sheet currently specifies 4 teams per topic, meaning 24 teams total). Assuming there are between 72 - 144 students in the class, there can easily be 3-6 people per team (respectively) to reach 24 teams. Having students choose their own team may be easier in person but, virtually, may be done by opening 24 breakout rooms and having students join whichever room they want until there is the correct number of students per team (this may however take a few extra minutes). Alternatively, the instructors/TA can set up randomly-selected teams if desired.

**Given an even number of teams per topic, groups can be paired with another team within their topic such that the thinking they did about the topic in their own group can more easily extend to the peer review.

Final class activity

- Introduce the activity in the lecture/slides
    a. Each pair of student teams will meet in a breakout room (if virtual) for 20 minutes to discuss the ethical considerations of each team's idea
    b. Discussion prompts are provided in the slides as well as a shared Google Doc for students to use while in breakout rooms
    c. After the 20 minutes is over, a random team-pair will be selected from a random three of the six social good topics to speak for up to 3 minutes about what they discussed
        - Each team-pair should therefore be instructed at the start of the activity to choose 1-2 representatives from their group and a few talking points before the breakout room section is over (potentially giving a five minute warning at the end to start this piece if they have not done so already)
        - It is optional to leave a few minutes for comments or questions from the rest of the class after each presentation; this may depend on the likelihood of participation and the time available
    d. The team-pairs that are not selected should post a summary of what they would have shared to the Piazza discussion board (have one person from the broader 2-group team add the post).

Supporting Materials

- See ***Appendix 5*** for slides with notes (the Powerpoint file is also shared in Google Drive)
- Google Drive includes links to the topic sign up Google Sheet and two Google Docs that students can use to review instructions and take collaborative notes while in breakout rooms
- Anonymous Authors (2020) Dataset Bias in Diagnostic AI Systems: Guidelines for Dataset Collection and Usage. https://doi.org/10.1145/1122445.1122456
- McLennan, S., et al. AI Ethics in Not a Panacea. The American Journal of Bioethics. 20(11): 20-22, https://doi.org/10.1080/15265161.2020.1819470

Evaluation

Plan: Evaluation for the module will consist of 2 parts, a graded component that will be reflected in the students' final score, and a post-class survey. The grading would take 3% of the whole class grade, taken from the 5% participation grade. A post-class survey is distributed to the students at the end of class to measure engagement rate and quality of student comprehension. Feedbacks in this post-class survey should be used to design further iterations of this module in the future.

Grading detail of total ethics grade (3% of whole class grade):

- **Discussion on explainability (0.5% - of the 3% above - to whole class grade):**
  Full grade will be given if it is completed (satisfies the minimum sentence requirement and done in time) and the provided answer is relevant to the topic at hand
- **Discussion on recommender systems (0.5% - of the 3% above - to whole class grade):**
  Full grade will be given if it is completed (satisfies the minimum sentence requirement and done in time) and the provided answer is relevant to the topic at hand
- **ML for Good innovation pitch (1% - of the 3% above - to whole class grade):**
  Full grade will be given if it is completed (satisfies the minimum sentence requirement and done in time) and the provided pitch is feasible and relevant to the topic at hand
    - Each person in the team should generally get the same grade
- **Peer review of innovation pitch (1% - of the 3% above - to whole class grade):**
  Full grade will be given if it is completed (satisfies the minimum sentence requirement and done in time) and the provided explanation makes clear that they've engaged with the other team's idea and directly connects to ethics concepts
    - Each person in the team should generally get the same grade

Post-class survey question points:

- Relevance of ethics materials to the topics being studied
- Opinion on whether the ethics materials are valuable
- Favorite components of the whole module (if any)
- Open-ended question on improvement and student's general comment

Supporting Materials

- See ***Appendix 6*** for the exact wording of questions

Limitations and Future Considerations

Limitations

Due to the limited time, scope and resources allotted in the design process of this module, there are some things we recognize as a known limitation. These limitations include:

- In general, we did not cover strategies and efforts already made in the AI/ML community to overcome the ethical questions being posted and focused more on things to be mindful of when conceptualizing and building ML models
- The notion of explainability, or lack thereof can apply to other topics beyond just decision trees - most notably deep neural network models
- Our coverage of recommender systems (via a podcast episode) might not touch deeply on why and how bias might happen in a technical sense, it focuses more on the high-level abstractions and moral questions. There are other literatures we found, including one by Chen, Dong, Wang, Feng, Wang and He (1) that captures ethical concepts at a more granular, technical level when applying recommender systems, e.g.:
    - Unfair advantage through a closed feedback loop for popular first movers in e-commerce like Amazon, making it harder to introduce variety and competition
    - The fact that people give feedback only if their experience is extremely good or extremely bad can skew the recommender system's output
- The activities and related instructions, especially the ones set up for the ML for Good activity, assumed the learners are willing participate and are equipped with ample communication ability to function in groups - we did not provide back up plans for on-ground challenges, as we think the instructor's judgment is better suited to tackle those impediments
- All the activities are designed with minimum disruption to the current class structure, which limited some opportunities for changing class order or homework content that may have created a better learning experience. For example, we had talked with the instructor about using a similar dataset in the decision tree and neural network homework to reflect directly on explainability, but decided against current implementation due to time and resource constraints in making that happen

1. Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2020). Bias and debias in recommender system: A survey and future directions. ArXiv:2010.03240 [Cs]. http://arxiv.org/abs/2010.03240

Possible future additions

There are other topics in the class that can be tied to ethics concepts but not included in the current ethics incorporations, like regularization and overfitting. More tie-ins for these other topics in the class can be added in the future, using the current module as reference. There is also some spare time for other additional activities or ethics content in the last two classes as the ML for Good activity does not take the entirety of the class time. One way to fill the remaining time is to loop in the existing ethics discussion built out for the 15-281 Artificial Intelligence: Representation and Problem Solving class, which may remain a lecture or be incorporated into the activity. If the latter, relevant questions to extend the activity may include:

- How these ML innovations would affect jobs?
- Can it be weaponized?
- How far might it intrude on people's privacy?
- Who is responsible if something goes wrong?

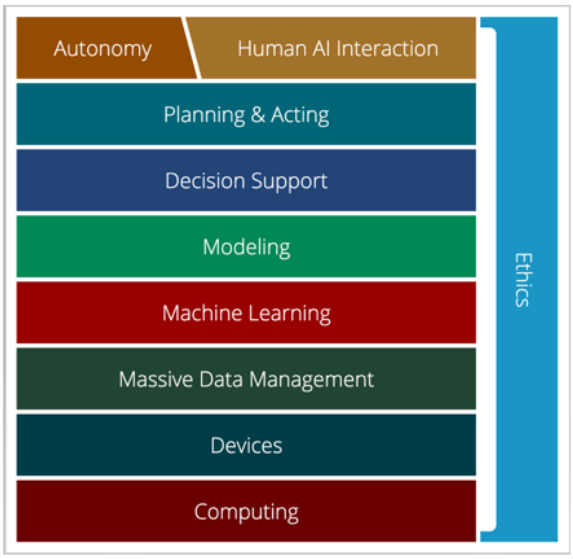
Appendix 1

Introductory slides and notes

Appendix 1: Introduction Slide Deck Additions

AI Stack for CMU AI

"AI must understand the human needs and it must make smart design decisions based on that understanding"

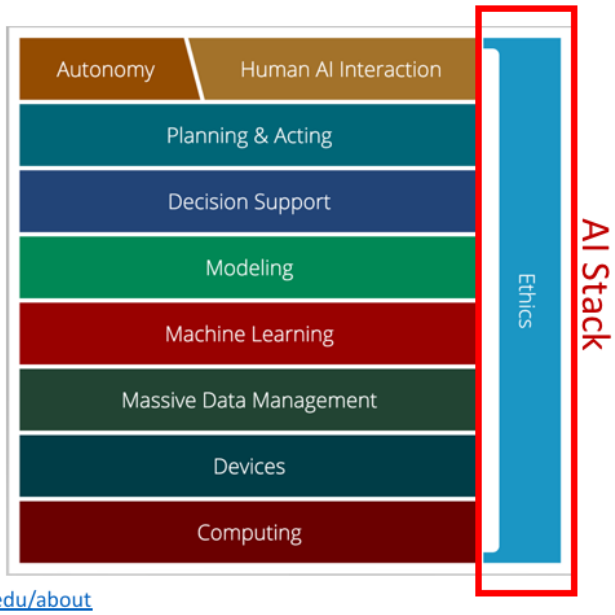


https://ai.cs.cmu.edu/about

[This - slide 6 in original lecture - is unchanged but is a lead in to the next slide] Proposed by Andrew Moore, CMU is using AI Stack to illustrate what they consider as AI. AI must understand the human needs and it must make smart design decisions based on that understanding. Despite the simple definition, though, AI isn't just one thing. It's a giant thing, built from technology blocks we call the AI Stack. At Carnegie Mellon, we view it as a toolbox — each block houses a set of technologies that scientists and researchers can reach for as they work on new initiatives. Expertise in all areas? Not required. Instead, we believe you can focus on one area and draw on other parts of the stack for help. Each block depends on the other for support. And AI endeavors that ignore parts of the stack won't succeed.

AI Stack for CMU AI

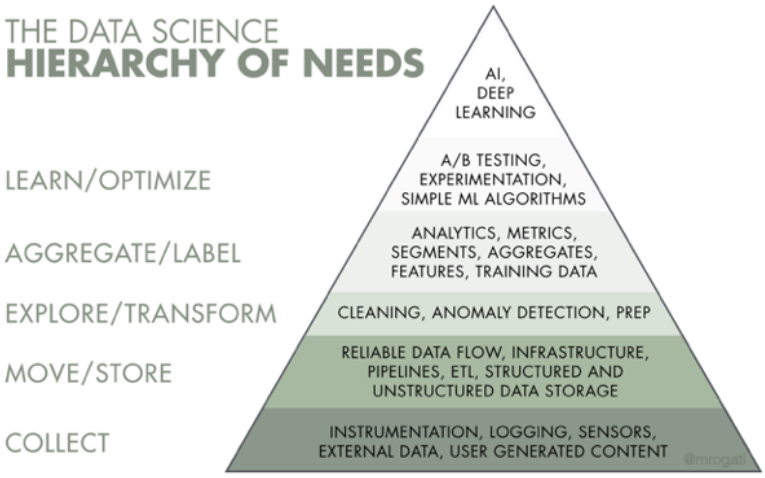
"AI must understand the human needs and it must make smart design decisions based on that understanding"



https://ai.cs.cmu.edu/about

[Slide 7] As you can see in the stack, the consideration and application of ethics spans all other categories. In a similar way, we will have ethics related discussions and homework questions throughout the semester signified with this visual cue.

The Importance of Input Quality



[To follow Slide 20 – The Machine Learning Framework; next two slides on dataset bias follow this one] Bringing our first bit of ethics into the conversation, let's look at this data science hierarchy of needs and the largest / most important consideration at the bottom – the specifics of data collection. I'll briefly mention two meaningful consideration when it comes to collecting and selecting data.

1. Considering Existing Patterns in the Data

Goal: Create an algorithm that uses resume information to help decide which engineers to hire.

Training data: Resumes of previously hired engineers.

What issue arose given this training data?

The algorithm systematically discriminated against women applying for technical roles, flagging:

- women's names
- explicit mentions of "women" (i.e. "women's rugby")
- names of women's colleges
- verbs less often used by males

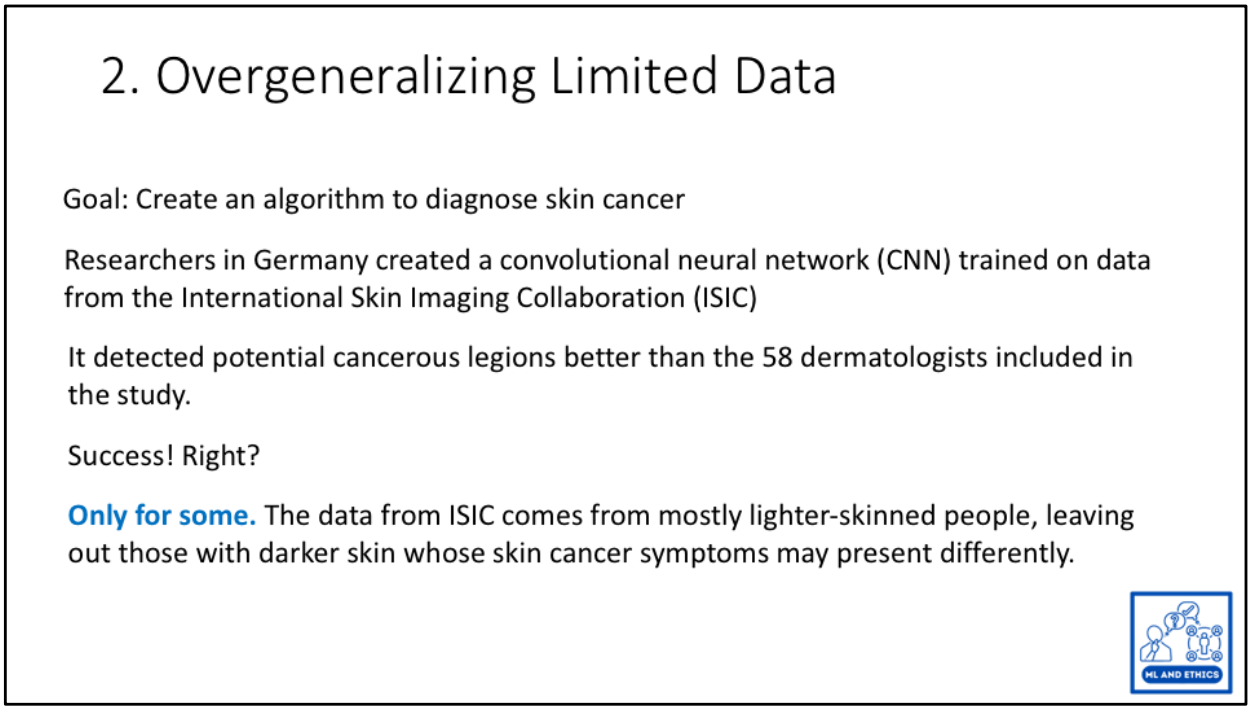
Why?

The first is considering existing patterns or biases that might be present in the training data that the algorithm may perpetuate or even exacerbate. One example that you may have heard of is Amazon's early attempts at creating a hiring algorithm back in 2014. The goal was to build a machine learning tool that could scan applicants resumes and help improve the identification and hiring of good engineering candidates. It was built off the training data of the resumes from previously hired engineers.

Any idea or thoughts about a big issue that arose from building a model based on this data? What ultimately happened was that the algorithm systematically discriminated against women applying for technical roles. It not only flagged resumes with female names and explicit mentions of "women" (like in sports team descriptions) but even the names of women's colleges and verbs used less often by males.

As you may have guessed, a driving force for this outcome was that an overwhelming majority of the previous resumes, representing those successfully hired, were from male candidates. Despite programmer's effort to reduce this bias, they were unable to do so fully and ultimately stopped using the algorithm.

Reference:
https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against



The second consideration is who the data is and is not representing and how this connects to where the algorithm is being applied.

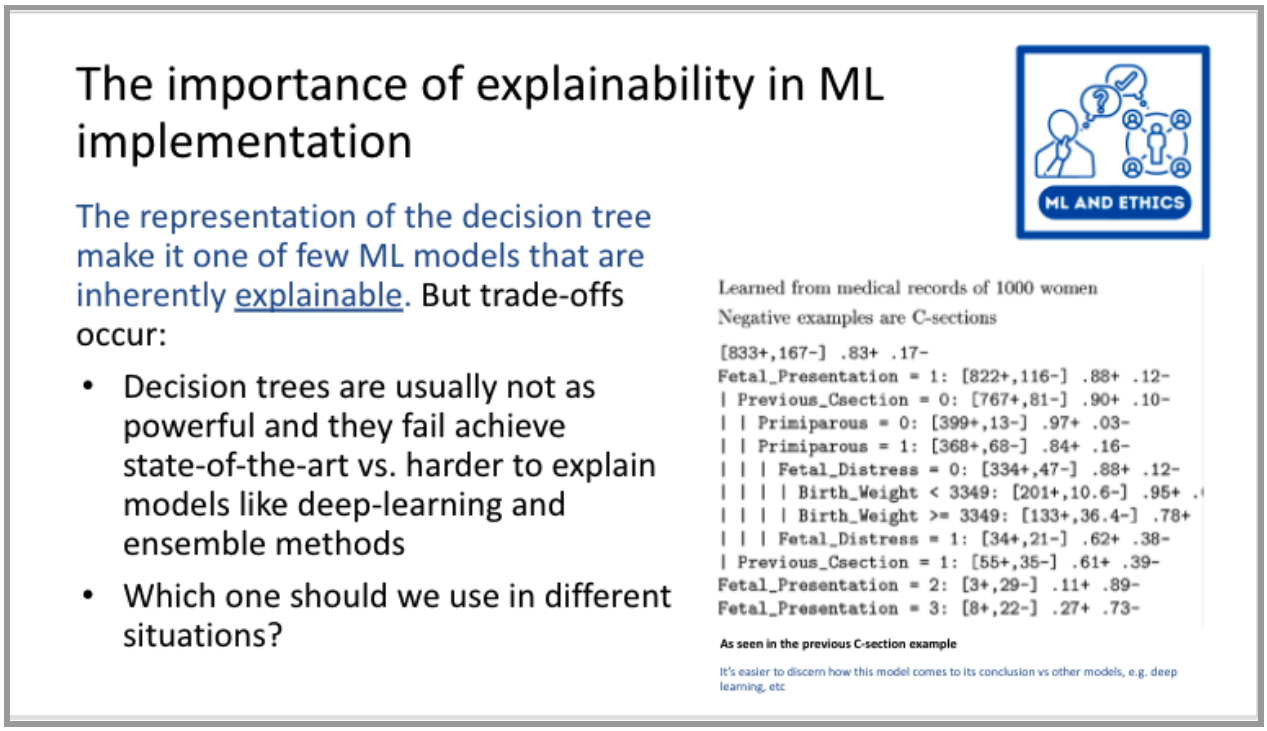
I already briefly mentioned medical diagnosis so let's talk more about the diagnosis of skin cancer. In 2018, researchers in Germany created a convolutional neural network (CNN) trained on data from the International Skin Imaging Collaboration (ISIC). The algorithm was ultimately able to detect potential cancerous legions better than the 58 dermatologists included in the study.

That was easy - these researchers already reached the goal, right?
Well, no. In fact, the skin imaging data used was mostly for lighter skinned patients even though early signs of skin cancer for those with darker skin can present differently. As algorithms like this are being rolled out and presumed to be used for all patients, these tools are more likely to then misdiagnose people of color with nonexistent skin cancers or miss them entirely. (Since the time of this research there has been a push to contribute medical images of a wider diversity of patients to the open-source ISIC)

Citation: https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/
https://pubmed.ncbi.nlm.nih.gov/29846502/

Appendix 2

Explainability slides and notes

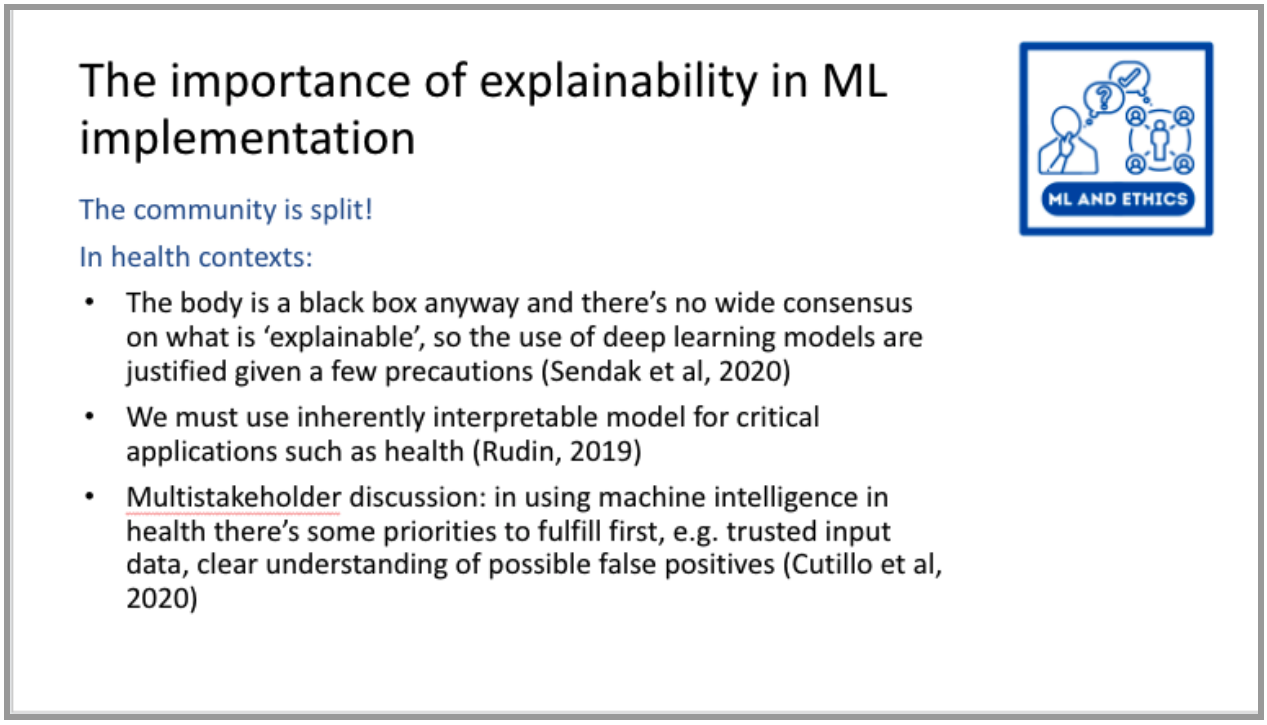


This points in this slide, including the inherent trade-off are taken from:

- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. MDPI AG. Retrieved from http://dx.doi.org/10.3390/e23010018

It might be good to revisit and ask the students to reflect explainability with example from previous slides (the C-section model is included here)

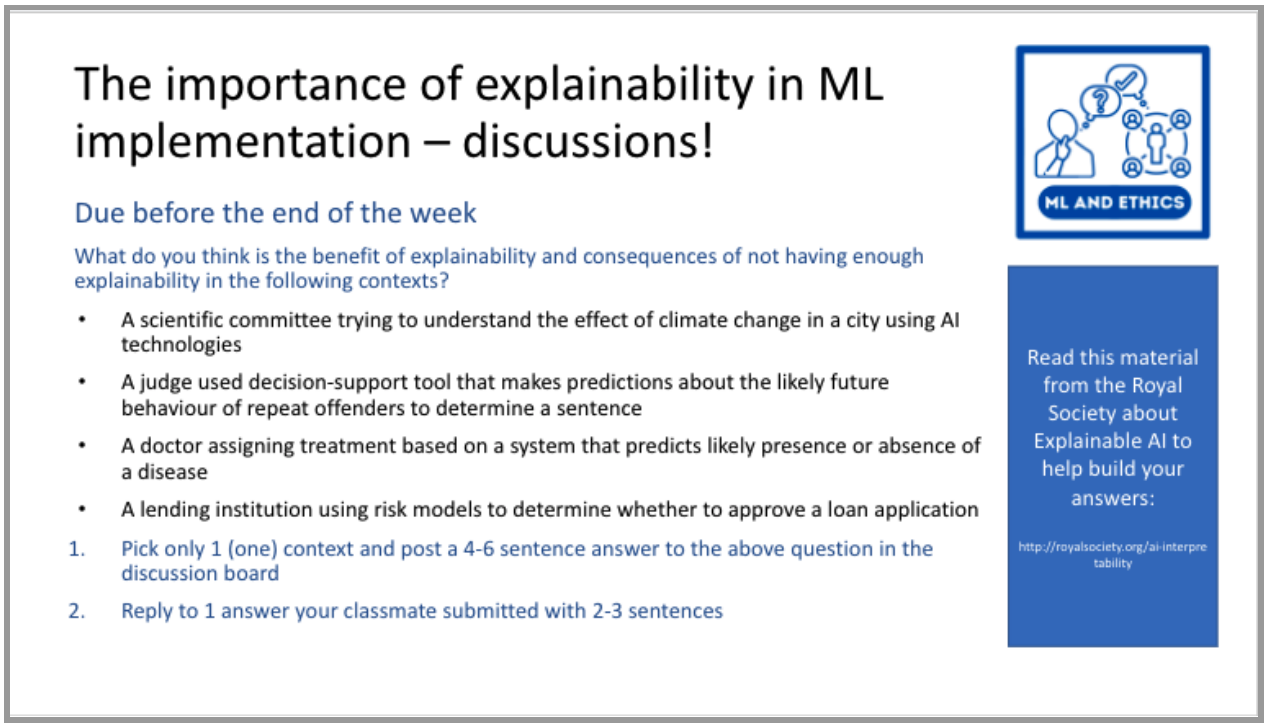
The last point is intended as a hanging question that can be used as a segue to the next slide



This slide is intended to show that the answer to the question on the previous slide is not that clear cut - in some sense state-of-the-art accuracy is needed, but in the context of change management and in terms of building trust in the medical community, explainable alternatives is clearly needed.

Points are taken from:

- Sendak, M., Elish, M., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., & O'Brien, C. (2019). "The human body is a black box": Supporting clinical decision-making with deep learning. ArXiv:1911.08089 [Cs]. http://arxiv.org/abs/1911.08089
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
- MI in Healthcare Workshop Working Group, Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., & Mandl, K. D. (2020). Machine intelligence in healthcare—Perspectives on trustworthiness, explainability, usability, and transparency. *Npj Digital Medicine*, *3*(1), 47. https://doi.org/10.1038/s41746-020-0254-2

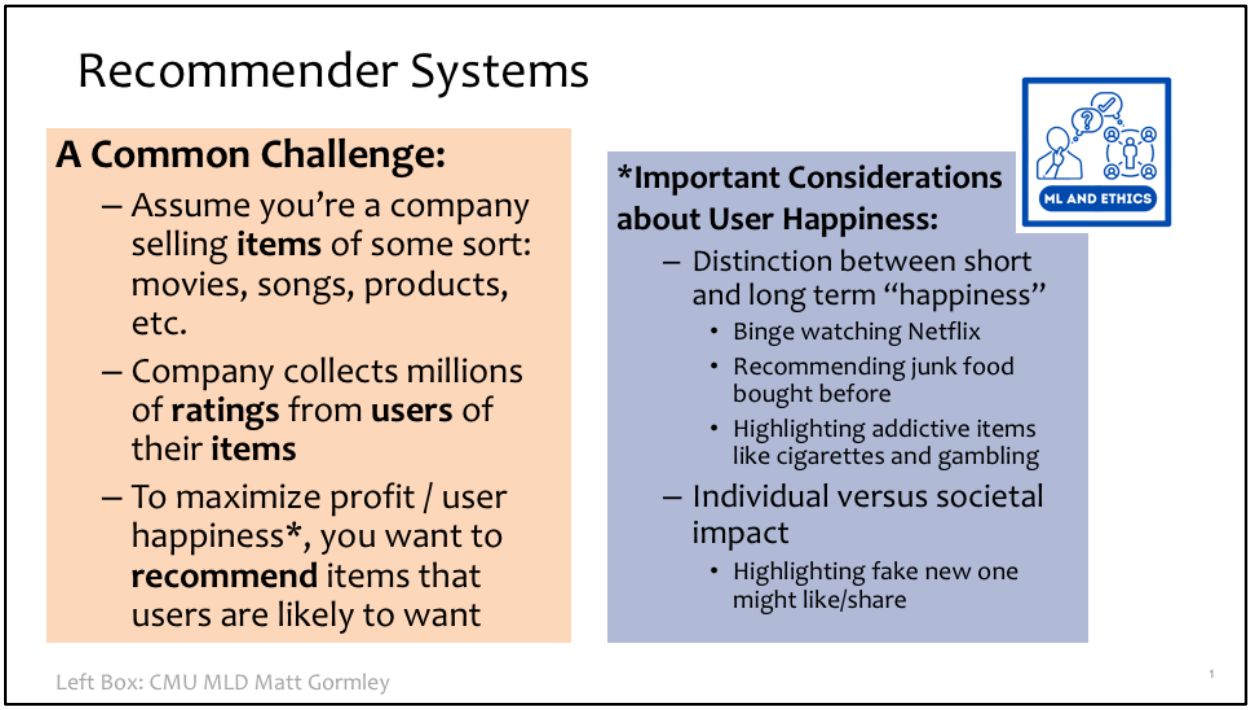


Begin with stating that despite the split opinion, most of the time, explainability is important - the instructor can touch with an example like:
“Imagine that the opinion of a cardiologist differs from the output of a clinical decision tool. What can happen if we have a semblance of explainability and what can happen if we don’t have it“

Then continue to ask the students to do the discussion board assignment. They can choose one topic out of four possible topics (science, criminal justice, health and finance) to reflect explainability on. Before making their reflection, ask them to read the material from Royal Society (The Royal Society (2019). Explainable AI: The Basics - Policy Briefing [White paper]. The Royal Society. Accessed April 27th, 2021 from http://royalsociety.org/ai-interpretability - from which the topics are taken out from), that can help them build a richer reflection.

Appendix 3

Recommender System slides and notes



[Left box is the same as original content; right box is added]
Description for right box:

- When you are recommended a new song you like or the next book you're excited to read, it may seem obvious that recommender systems help to maximize your happiness. However, is this always the case? One potential way to break this down is thinking about the distinction between short and long-term happiness. Finding that next great song or book might bring you happiness on both scales but there are also other instances that might only right the short time side and even backfire long term.
- One example you may or may not relate to is binge watching a show on Netflix. In the moment it is often exhilarating but when your ML homework is due a few days later, you may look back on it with less happiness. This can also be true when recommender systems use previously displayed preferences to frequently highlight things like junk food, cigarettes or online gambling to a user. For example, for someone who recently decided to try to eat more healthy food or quit smoking, using a platform with such a recommender system could be triggering, disempowering and even influential enough to veer them away from their personal goals.
- Beyond whether recommender systems lead to individual happiness, it is also important to consider societal impact and well being. One can look no further than political influence and polarization on platforms like Facebook and Twitter. Algorithms behind sorting and prioritizing news feeds have also influenced vaccine behavior, which in this covid time is of renewed consequence.

Recommender System Assignment

1. Listen to episode of *Towards Data Science* podcast on Ethical Problems with Recommender Systems
2. With 4-6 sentences each, answer **2 of the 5 question** prompts on the Piazza discussion board
3. Reply to **at least two** of your fellow classmate's responses with 2-4 sentences

[Final slide of deck] Description of ethics component of HW assignment; there is additional details (including the question prompts) in the HW document itself.

Appendix 4

Recommender System homework description

HW9- Recommender Systems

1. Listen to episode of *Towards Data Science* podcast on Ethical Problems with Recommender Systems
    - Listen or read the transcript here: https://towardsdatascience.com/ethical-problems-with-recommender-systeems-398198b5a4d2
    - Optional additional reading: the research paper from the podcast guest Silvia Milano https://link.springer.com/article/10.1007/s00146-020-00950-y

2. With 4-6 sentences each, answer **2 of the 5 question** prompts below on the Piazza discussion board
    - Should users have rights when it comes to recommender systems (like they might with data privacy or collection)? Does this answer change depending on if people have fixed or malleable preferences?
    - Jeremie (the host) said: "Before I ever use Twitter, my political views were some set of beliefs. And then after I use Twitter, my political views were a different set of beliefs. I changed as a person from that interaction." Have you had an experience where it seemed that a recommender system noticeably influenced your beliefs, decisions or actions?
    - Are there certain areas, like jurors deciding guilt or innocence beyond a reasonable doubt, that should not be quantified?
    - What are your thoughts on shifting from a user centered to multi stakeholder approach? (Or is there a better third way?)
    - What characteristics would you prioritize to construct an ethical recommender system?

3. Reply to **at least two** of your fellow classmate's responses with 2-4 sentences

    - Should users have rights when it comes to recommender systems (like they might with data privacy or collection)? Does this answer change depending on if people have fixed or malleable preferences?
    - Jeremie (the host) said: "Before I ever used Twitter, my political views were some set of beliefs. And then after I used Twitter, my political views were a different set of beliefs. I changed as a person from that interaction." Have you had an experience where it seemed that a recommender system noticeably influenced your beliefs, decisions or actions?
    - Are there certain areas, like jurors deciding guilt or innocence beyond a reasonable doubt, that should not be quantified?
    - What are your thoughts on shifting from a user centered to multi stakeholder approach? (Or is there a better third way?)
    - What characteristics would you prioritize to construct an ethical recommender system?

Appendix 5

ML for Good slides and notes

Machine Learning for Good: Class 1

How can machine learning be used for social good & what ethical considerations are there?

Your Turn to be the Inventor!

Healthcare

Privacy

Financial Health

Sustainability

Education

Civic Engagement

Breakout Room (20 min)

1) Agree upon and sign up for one of these six categories

2) Brainstorm machine learning applications for social good in your selected category

3) Pick one idea and write a 6-8 sentence summary, including a brief plan for data collection and model implementation

*This will be shared with the class through Piazza

Breakout rooms for the next 20 minutes
Randomly assign the number of teams that allows for 24 teams (such that 4 teams can sign up per category)
(*might need to shift if there is less than 72 students or more than 144)
Sign up with team names under category: up to 4 teams per category
*Send link in chat to google sheet like this one:
https://docs.google.com/spreadsheets/d/1QndIhrAb05cDCnCRSUwUiSIKacL0nBPmMo8QuyW_UFw/edit#gid=0

*Let students know that the idea does not necessarily need to be something that they can fully build themselves
Images:
https://www.forbes.com/sites/bernardmarr/2019/11/01/the-9-biggest-technology-trends-that-will-transform-medicine-and-healthcare-in-2020/
https://npl971975.wordpress.com/2019/09/05/collection-five-rules-to-improve-your-financial-health/
https://fitizen.co.in/how-to-use-technology-to-build-environmental-sustainability/
https://www.mcall.com/news/pennsylvania/mc-pa-girl-graduates-college-before-high-school-0509-20170509-story.html
http://www.longbeach.gov/cityclerk/services/civic-engagement/

Create and Share Your Short Pitch

Assignment

1) Have **one member** of your team post a short description of your idea to the Piazza discussion board

- 6-8 sentences, including brief plan for data collection and model implementation
- Tag your selected topic area
- Include names of your teammates
- Due:

2) Review the idea of the other group you are assigned

- Think about ethical considerations of the idea
- Be prepared to discuss this topic with the other team during our next class
- This is not a critique, but rather, the next step in strengthening your ideas!
- Due: [date/time of next class]

Image:
https://hbr.org/2003/09/how-to-pitch-a-brilliant-idea

Final Class

Machine Learning for Good: Class 2

How can machine learning be used for social good & what ethical considerations are there?

Data collection and dataset bias - hierarchy of needs, amazon recruiting, skin cancer
Explainability - judge (likely future behaviour of repeat offenders), lending decision (examples of when explainability may be particularly important)
Recommender Systems - toward data science podcast, example of frequently recommending addictive items

Images:
https://wklconsultancy.nl/recognize-good-recruiter/
https://www.bicycling.com/training/g20044455/5-signs-of-skin-cancer-that-are-easy-to-overlook/
https://www.abajournal.com/web/article/judges-personal-relationships-formal-opinion-488
https://www.smartaboutmoney.org/Courses/Money-Basics/Credit-and-Debt/How-Do-Lenders-Make-Money
https://www.womenwork.org/different-types-of-addiction/

Peer Review Discussion (20 minutes)

Bias: An unintended or potential harmful property of data or its implementation.

Ethics Discussion Prompts

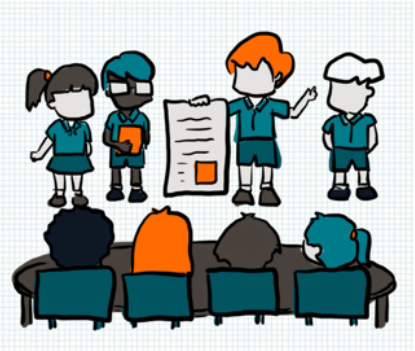
- What biases may arise based on the plan for collecting data (be it by the team or using existing datasets)? For example, consider:
  - representation
  - any discrepancies in data quality due to measurement differences across populations or data sets
  - historical biases already embedded
- Beyond the exact context of the training data, for which other situations can the model be generalized, if any? What are the risks of doing so incorrectly?
- Who will the model be directly affecting the most and what could go wrong for this population? For example, consider:
  - How could the team learn more about this population before and during model building?
  - Is there a potential for the model's impact to change over time? If so, how might monitoring play a role?
  - What should an end-user do if an issue arises? Who is responsible?
- What might be the societal implications? For example, consider:
  - How jobs may be affected now or in the future
  - How far it may intrude on people's privacy
  - Could it be misused intentionally or unintentionally

Groups can consider topics beyond these questions; these are meant as a starting point / guide rails for the discussion – and should feel free to start from anywhere (as opposed to from top to bottom)

Could include the contents of this slide in a pdf or piazza post for students to refer to while in breakout rooms

Citation (and could provide as optional reading between classes):
http://web.mit.edu/juliev/www/CHIL_paper_bias.pdf
https://www.tandfonline.com/doi/epub/10.1080/15265161.2020.1819470?needAccess=true

What Did You Discuss?



- One pair of teams from each of the 6 categories will be asked to share the key points of your ethics discussion for 3 minutes.
- While in your breakout rooms, choose 1-2 representatives and just one of the two ideas to highlight.
- The teams that are not selected should post a summary of what they would have shared to the Piazza discussion board (have one person from the broader 2-group team add the post).
  - Due:

Image: https://www.learner.org/wp-content/uploads/2020/05/two-bit-circus-lesson-plans-unit-elementary-school-engineering-towers-group-presentation-scaled.jpg

Appendix 6

Post-class survey document

Questions

Relevance

How relevant do you think the ethics material (those signified by the Ethics and ML logo) is to the class topic being discussed?
[Likert scale]
1 (Not relevant at all) 2 (Not so relevant) 3 (Quite relevant) 4 (Highly relevant, essential)

Explain the reason for your opinion
[Paragraph answer]

Value

Do you think the ethics material infused in this class is valuable to you as a CS graduate?
[Likert scale]
1 (Not valuable at all) 2 (Not so valuable) 3 (Quite valuable) 4 (Highly valuable)

Explain the reason for your opinion
[Paragraph answer]

Engagement

What ethics activity do you like the most?
[Multiple choice]

a. The introduction lecture and the concept of dataset bias
b. The discussion around explainability
c. The discussion around recommender systems
d. The final class activity

Explain the reason for your choice
[Paragraph answer]

General Comments

What can be improved if we want to continue this part of the class moving forward, and do you have any other comments on the whole activity?
(Open-ended questions)
[Paragraph answer]