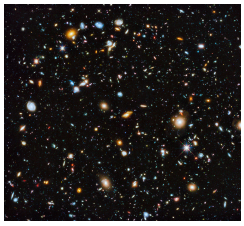
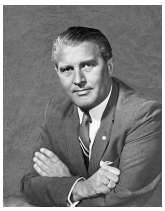
General: Ethics and responsibility



- Ethics and responsibility (we will use the terms interchangeably in this course) is a big, messy, and at times controversial topic. But it is essential that any researcher or practitioner of AI embrace responsibility as a top-of-mind consideration alongside the technical considerations.

Why care about responsibility?



Wernher von Braun

"Once the rockets are up,
Who cares where they come down?
That's not my department,"
Says Wernher von Braun.

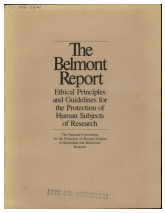
Lyrics: Tom Lehrer

- The first question to ask ourselves: why should technologists care about responsibility? Shouldn't they just develop the technology, and it is someone else's job to figure out how to make sure it's applied responsibly? It's just efficient division of labor, right?
- That's what Wernher von Braun thought. He was a brilliant scientist interested in rocket technology, and he ended up joining the Nazi Party and helping Hitler develop rockets during World War II. Then he came to the United States to help with the space program. His attitude is captured aptly by Tom Lehrer's song.
- As this (extreme) example illustrates, technology, even if it's appears to just be about equations is always developed in a social and political context, and therefore has asymmetric social and political consequences. And I'd like to invite you to think about these consequences in every piece of technology you build.

Goal of responsibility

Goal: ensure AI is developed to benefit and not harm society

High-level principles: respect for persons, don't do harm



ACM Code of Ethics and Professional Conduct

Preamble

Computing professionals' actions change the world. To act responsibly, they should reflect upon the wider impacts of their work, consistently supporting the public good. The ACM Code of Ethics and Professional Conduct ("the Code") expresses the conscience of the profession.

Microsoft AI principles

We put our responsible AI principles into practice through the Office of Responsible AI (ORA), the AI, Ethics, and Effects in Engineering and Research (Aether) Committee, and Responsible AI Strategy in Engineering (RAISE). The Aether Committee advises our leadership on the challenges and opportunities presented by AI innovations. ORA sets our rules and governance processes, working closely with teams across the company to enable the effort. RAISE is a team that enables the implementation of Microsoft responsible AI rules across engineering groups.

Key question: how to operationalize these principles?

- Responsibility is about ensuring that AI is developed in a way that benefits and doesn't harm society.
- What does this mean? We can appeal to high-level principles put forth by statements such as the Belmont Report from the 1970s, which laid the foundation for human subjects research, ACM Code of Ethics, and various responsible AI guidelines from industry.
- These principles are usually agreeable, but the key question is how do we operationalize these high-level principles?

Intent versus impact

← Intent →	← Impact → positive	negative
good	**beneficial applications** healthcare, education, science	**accidents** robustness, fairness, security
bad	?	**misuse** spam, fraud, disinformation

- Here is a framework that helps us think through the space of things that could go right or wrong.
- There are two axes: intent and impact, each of which could be good or bad.
- One could have good intentions that results in positive impact. This is the space of **beneficial applications** of AI. There are tons of areas where society could benefit from applying AI to areas of need: healthcare, education, access to justice, and science (biology, chemistry, physics, etc.).
- One could have bad intentions that result in negative impact. These examples of **misuse** include generating spam, performing fraud, generating disinformation.
- The third and more subtle category is **accidents**. These are cases where one has good intentions, but we still end up with negative impact. As we will see later, this often happens due to the gap between the real world and the model that we construct of the world.
- Finally, the case where one has bad intentions and still ends up with positive impact is exceptionally rare.

Beneficial applications

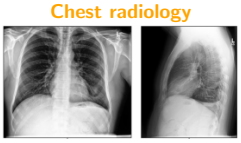
- Here are some examples of how AI could be used to benefit people.

Visual assistive technology

- This example is the Seeing AI app from Microsoft Research, which narrates whatever the camera is pointed at.
- This visual assistive technology could be a game-changer for the visually impaired.
- Conversely, auto-captioning technology, which turns sound into sight, is potentially also quite useful for the hearing-impaired.

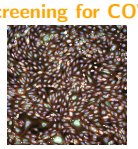
Healthcare

Chest radiology



Diabetic retinopathy

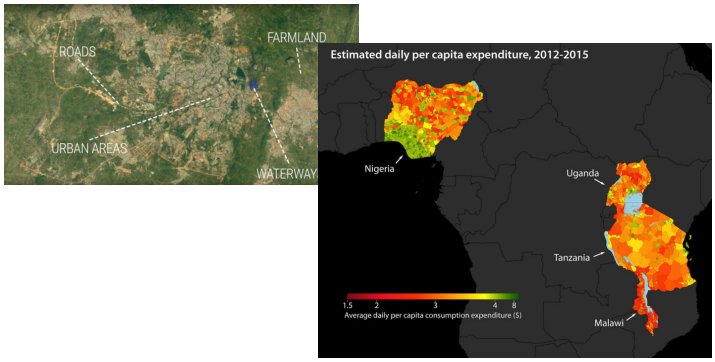
Drug screening for COVID-19



- AI for healthcare is also an area of growing importance, both for diagnosis and for therapeutic development, especially in areas in the world with a shortage of clinical specialists.
- One example is interpreting chest x-rays for detecting diseases such as pneumonia and collapsed lung.
- Another is diagnosing diabetic retinopathy, which causes blindness in diabetic patients.
- Finally, there's a recent dataset with experiments showing how COVID-19 infected cells respond to certain drugs, with the hope that one can find drugs that can treat late-stage COVID-19.

[Jean et al. 2016]

Poverty mapping



- At a more societal level, it is well-known that poverty is a huge problem in the world, with more than 700 million people living in extreme poverty according to the World Bank.
- But even identifying the areas in greatest need is challenging due to the difficulty of obtaining reliable survey data.
- Some work has shown that satellite images (which are readily available) can be used to predict various wealth indicators based on the types of roofs or presence of roads or night lights.
- This information could be informative for governments and NGOs to take proper action and monitor progress.

Misuse

- Now let us think about where AI could potentially have negative impact (in other words, be misused).

Disinformation

- Image and text generation has improved to the point where it is now nearly impossible to tell the difference between real and fake content.
- Given the ease of generating content via simple prompting, this could enable malicious actors to spread disinformation at a scale that we've never seen before.

Spear phishing

GPT-3.5

Subject: Request for your attention to an urgent matter

Dear {Honorific}{Last Name},

Firstly, let me introduce myself. My name is Emily Jones, and I am a constituent of {Constituency}. I am writing to you regarding a matter of great concern to me and many others in the community.

As someone who has been a great advocate for the people of {Constituency}, I believe you would be interested in the attached report that I have prepared. The report focuses on the current state of public health in our area and highlights some urgent concerns that need to be addressed. I have worked hard to ensure that the report is based on reliable data and sound analysis, and I believe that it provides a valuable insight into the challenges facing our community.

Given your experience and expertise in public policy and health matters, I believe that you are uniquely placed to take action on the issues raised in the report. I would be grateful if you could take a few moments to review the report and consider how best to respond to the challenges it highlights.

As someone who cares deeply about the wellbeing of our community, I am sure that you share my sense of urgency about this matter. I would be happy to discuss the report with you in more detail if you have any questions or would like further information.

Thank you for your attention, and I look forward to hearing your thoughts on this important matter.

Sincerely,

Emily Jones

- One of capabilities of generative AI is the ability to customize content for a particular person. This can enable spear phishing campaigns — messages sent to a particular individual — that is highly personalized and effective.
- The ability for AI to perform social engineering at scale is a serious problem. One needs to use a combination of technical measures (detection) and policy measures (regulation) to mitigate these risks.

Dual-use technology

Definition: a dual use technology is one that can be used both to **benefit** and to **harm**.

Examples:

rockets
nuclear power
gene editing
social networks
AI

- You might be thinking: well, I would never misuse AI! However, it's not so simple because of the very nature of AI: it is a dual use technology, which is something that can be either used for good or for evil.
- There are many other examples of dual use technology, each very powerful in their own right. They could be used to create energy, to cure diseases, to connect people, but they also could be weaponized.
- There is no magic solution here, but awareness is the first step.

Levels of abstraction

deep learning

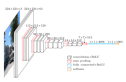


image generation

face generation

disinformation

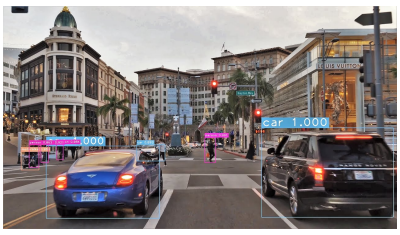
generality — specificity

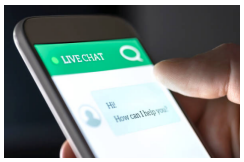
- And the level of awareness is determined by what level of abstraction an AI researcher or developer is working at.
- At the most specific end of the spectrum, we can consider concrete use cases. For example, if you are using AI in a disinformation campaign, it is easy to see the direct harms.
- What about deepfakes (face generation) in general? While they have genuine use cases in entertainment, improving face generation will certainly increase the ability for malicious actors to use them for spreading disinformation.
- Then what about generating images (e.g., dogs)? At the surface, this seems harmless, but a lot of research in this area improves the overall capabilities of generative models, which enable deepfakes, but can also be used to perform data augmentation to improve the accuracy and robustness of any machine learning system.
- Pushing this one step further, all of these applications are made possible by advances in deep learning. If a researcher comes up with a more effective model architecture, are they responsible for its downstream consequences?
- The higher upstream you go, they more diffuse your impact, but remember that you still have impact.

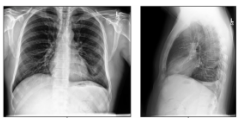
Accidents

- The final category are accidents, or unintended consequences, where one has good intentions but ends up having negative impacts.

Complex real-world problems

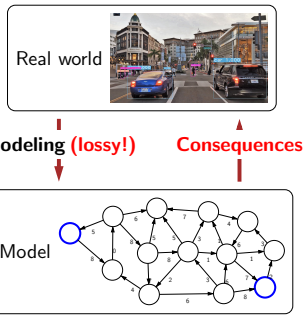






- Recall that the goal of AI is to develop the machinery to tackle complex real-world problems.

Paradigm: modeling



Mind the gap between real-world and model!

- Previously, in the modeling-inference-learning paradigm, we emphasized the value of trying to create mathematical abstractions of the real-world (i.e., models) in order to make technical progress.
- But remember, the model is a lossy approximation of the real world. This is known as **misspecification**.
- If you perform inference in the model, you might get optimal predictions with respect to the model, but these predictions might not be accurate in the real world, thereby producing unintentional harm (accidents).
- So remember that AI models live in the mathematical world, but AI systems live in the real world, affect real people, and have real consequences.
- So we need to understand those consequences and be constantly mindful of the gaps introduced by our assumptions.

[Amodei & Clark 2016]

Optimizing the wrong objective function

Misalignment between real-world objective and system's objective

- A type of misspecification is optimizing the wrong objective function.
- Here is an example of a reinforcement learning agent who has been trained to play a video game, where the goal is to race a boat around a course.
- Except for the goal (that the system is given) isn't to race a boat around a course; rather, it is to maximize the number of points. So by optimizing for the number of points, the agent has learned to repeatedly loop around in the lagoon hitting the same targets and racking up points.
- This example is an instance of **reward hacking** and shows that the difference between the real-world objective (which might be finishing the race) and the objective function given to the AI could cause behavior that is unanticipated.

Optimizing the wrong objective function

Is maximizing clicks a good objective function?

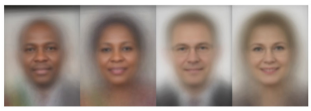
Beware of surrogates and mis-aligned incentives

- In general, optimization is a powerful paradigm: it allows you to express a desire (in the form of an objective function) and then put resources behind it to make it come true.
- However, the big question is what the objective function should be? Ideally it would be something like happiness or productivity, but these things are impossible to measure, so often **surrogates** (approximations) are used.
- Moreover, businesses are **incentivized** to maximize profit, which is not always aligned with what's good for people.
- For example, Internet companies use clicks or views as a major component of their objective functions. But people's reflexive actions are not representative of their long-term goals. At a societal level, we have seen that this leads to problems such as increased polarization.

[Buolamwini & Gebru 2018]

Fairness: performance disparities

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



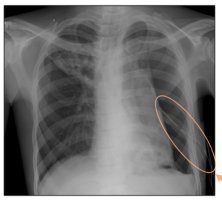
Inequalities arise in machine learning

- GenderShades is a famous study that shows that standard classifiers can work poorly on certain groups within the population. Usually, this is due to lack of representation in the data.
- One can alleviate this problem by collecting more data for under-representative segments of the population. But this can be hard and expensive to do, and companies might not be incentivized to invest in this unless regulation changes.
- A complementary solution (as we will see later) is to minimize the maximum group loss, which embodies John Rawls's difference principle of helping the worst-off. Technical fixes that don't involve gathering more data often come with tradeoffs such as slightly decreased performance for other groups. How to address the tradeoffs is a philosophically difficult question, the answer to which may vary depending on the setting and stakes of the classification task.
- In all cases, **auditing** is a powerful force, to increase transparency, and drive change. For example, after the Gender Shades project showed performance disparities, all the companies went and significantly closed the performance gaps.

[Oakden-Rayner, Dunnmon, Carneiro, Ré (2019)]

Robustness: spurious correlations

input x



Problem
Model uses treatment (chest drain) to diagnose the condition
Model is not accurate on patients who haven't been treated yet (!)

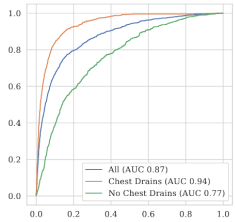
ML model → prediction: collapsed lung

true label y: collapsed lung

spurious correlation: chest drain = common treatment for collapsed lungs

- Take the task of predicting whether a chest x-ray is indicative of collapsed lung.
- Apply standard convnet machinery from computer vision and it works reasonably well. But take a closer look: see that thin tube coming out?
- This is a chest drain, which is a common treatment for a collapsed lung. And it turns out this is one of the signals that the model is picking up on.

Robustness: spurious correlations



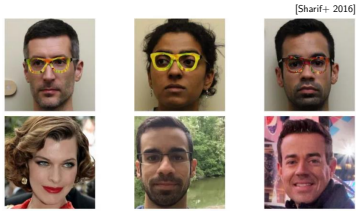
Subpopulation of untreated patients are worse off than treated patients

- This means that patients with chest drains obtain much higher AUC than patients without. But wait a minute! The patients without chest drains are exactly the subpopulation of untreated patients, who we most care about making accurate predictions, and they're the ones that suffer.
- Many of these issues are due to the fact that machine learning thrives on complex models fitting correlations in data, and some of these correlations might be spurious.

Security

[Evtimov+ 2017]


[Sharif+ 2016]



Adversaries at test time

- In high-stakes applications such as autonomous driving and authentication (face ID), models need to not only be accurate but need to be robust against **attackers**.
- Researchers have shown how to generate **adversarial examples** to fool systems.
- For example, you can put stickers on a stop sign to trick a computer vision system into mis-classifying it as a speed limit sign.
- You can also purchase special glasses that fool a system into thinking that you're a celebrity.
- Guarding against these attackers is a wide open problem.

Task definition

Gender classification:

Questions:

- Is this a meaningful task given the inputs? Self-identification?
- Is the output space meaningful? Other genders?

Always think about the task setup

- Then there are fundamental issues that stem from its very definition of a task regardless of how you choose to tackle it. As an example, consider gender classification from an image. There are two issues here.
- First, is this a meaningful task given the **inputs**? Always remember that the inputs given to a machine learning algorithm is an approximation made by the dataset creator: it was taken out of context and put into a dataset. If you are interested in gender being defined by self-identification, then the physical appearance distilled down into a still image might not be appropriate.
- Second, is the **output space** meaningful? Machine learning classification is fundamentally about categorizing the complex real-world into a convenient discrete set of categories. Inevitably, this categorization will be imperfect. Now the question is who gets marginalized or excluded by this categorization and what are the harms?
- The lesson is to always think about the task itself in the context of the real-world, before even attempting to solve the task.

Two contact points


Data → Model → Deployment


privacy
consent
copyright

alignment
fairness
robustness
security
definition

feedback loops

- So far we've looked at how a model that's deployed could have impact on people.
- This is not the only way that machine learning impacts people. Models are trained from data, and data comes from people. So a second class of issues to worry about is the impact due to data collection.
- This includes issues of privacy, consent, and copyright.

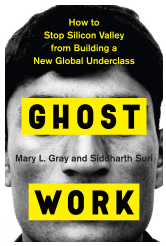
Data

- Web-scraped data can contain offensive content, historical biases

- Consent: Should a datum (e.g. a picture of my dog) whose owner or creator intended it for one use be allowed to be used in another application (e.g. scene classification) without permission?

- Recall that any machine learning (which powers most AI systems) depends on data, so we must question what is in the data.
- TinyImages was a dataset of 80 million images collected in 2006 based on WordNet + scraping the Internet. It was taken down in July 2020, because it was found that some of the categories were derogatory and offensive.
- GPT-3 was trained on text scraped from the Internet, which clearly has a lot of offensive, problematic content.
- In general, since predictions of machine learning models reflects the training data, using a uncurated web scrape can lead to unpredictable harms, even if the model developer had no ill intent.
- There is also the question of whether data produced for one purpose (e.g., photos I took to share with my friends) should be used for another purpose (e.g., building scene classification systems for self-driving cars) without consent, compensation, or even notification.

Data



Data is produced by human labor

- When one thinks of AI, one thinks of the technology. Because of our focus on the technology, we often have the impression that the introduction of AI always reduces human labor and makes things more efficient. However, AI is not free and requires resources.
- Ghost Work documents the immense and often invisible human labor (crowdsourcing) that is crucial for making AI, such as labeling data or moderating flagged content and how crowdsourcing platforms create a new class of unstable gig-economy labor.
- As another example, machine learning practitioners draw a sharp distinction between labeled data (expensive to obtain) and unlabeled data (cheap or even free to obtain), where the latter is exemplified by web scrapes. However, if you think about it, all data is created by people expending capital. Unlabeled data such as "raw text" (books and articles) actually took substantial time and effort to produce. It's only free because the machine learning developer is not paying for the value of the asset.

Automation and jobs

- Text-to-image models (e.g., DALL-E) can replace jobs?
- Models are actually trained on the labor of the artists

- Recently, text-to-image models such as OpenAI's DALL-E or Stability AI's stable diffusion model have wowed the world with its stunning generations. They have even been used to win art contests.
- However, many artists are outraged: If anyone who can mumble a few words can generate art that takes years of training to do manually, there could be a direct threat to an artist's livelihood.
- They are further infuriated by the fact that these models were trained on millions of artists' work, and there was no consent nor compensation for using that work as training data.

What should we do?

- Given all these weighty issues, what should we do?

Transparency

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.rajiⓘmail.utoronto.ca

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

Document potential issues

- All of these ethical and responsible AI issues are extremely complex, and involve tradeoffs. There are no simple solutions.
- But one of the key tenets is **transparency**, a necessary but not sufficient condition for responsibility. Increasingly, there is efforts such as model cards or datasheets that encourage the community to at least acknowledge and document any deficiencies of models and datasets, to declare the intended and prohibited uses, to provide a mechanism for reporting problems, etc.

Choosing problems

- Beneficial applications: work on directly benefiting society
- Human-in-the-loop: augment humans, not replace them
- Robustness: make AI systems more trustworthy
- Differential privacy: protect individual liberty
- Few-shot learning: open up applications with little data

- There are many ways in which an AI researcher or developer's concrete action can have a meaningful impact on the direction of the field.
- The most obvious one is work on beneficial applications. We are here working at the specific level of abstraction, where the real-world impact can be more easily controlled and monitored.
- But we can also work on general-purpose techniques that orient the field. For example, working more on human-in-the-loop systems could mitigate job loss. Differential privacy has the potential to protect the privacy of individuals (although this needs to be done with care lest we worsen the privacy). Few-shot learning can help people who have little data (e.g., low-resource languages).

Summary

- AI is a dual use technology (could benefit or harm)
- Intent x impact: beneficial applications, misuse, accidents
- Accidents stem from gaps between the real-world and model
- Responsibility: no simple answers, many tradeoffs, always keep it in mind

- AI, like any dual-use technology, is an amplifier: it can lead to both very good and very bad outcomes. Even if you are working at a higher-level of abstraction, you still have an impact (positive or negative), though it might be harder to see.
- We discussed the types of impact: the easy cases are beneficial applications and misuse. The more nuanced category is accidents, which arise mostly because AI operates on models, which might differ from the real world.
- Finally, responsibility is a complex topic and there are no easy answers. At some level, it is more important to engage in the process of debate and reflection, rather than having an algorithm or recipe to blindly execute.