CS 232: Artificial Intelligence

Spring 2024

Prof. Carolyn Anderson
Wellesley College

Reminders

- HW8 extended to today due to MarMon
- Lepei has help hours today and Thursday
- My Friday help hours are canceled but I will have help hours on Thursday from 10-10:30am
- Also feel free to email me for individual appts or with questions!
- HW 9 is a final project checkpoint

Bonus Late Day Opportunity

AI for Wireless and Wireless for AI: A Tale of Two AIs

Francesco Restuccia
Northeastern University

4-5pm
April 23rd

Bonus Late Day Opportunity

It's Wasmtime: Secure Isolation in Practice with WebAssembly

Chris Fallin
Fastly

9:55-11:10am

Thursday, April 22nd

SCI H401

What do we want the world to be like?

Vocabulary time!

Epistemic: related to knowledge. Epistemic questions are about what is true, what is known, or what is possible.

You can have a dessert (dessert exists).

Deontic: related to duty or to desire. Deontic questions are about what should or ought to be according to some set of obligations, desires, or norms.

You can have a dessert (you are allowed to).

Normative: related to an evaluative standard. Normative statements say how things *should* be, not how they are.

Evaluating AI Harms

Evaluating the potential harm of an AI system is a **normative question**. To judge whether a system is harmful, we need to decide what behavior is desirable.

What are some normative beliefs you hold about AI?

In other words, what are some things you think *should* be true about AI systems?

AI shouldn't: abuse data access
incorperate bias
mislead people
align w/ international law

Responsive to different needs of different people

How to accommodate different users in an unbiased way?

Data diversity & developer concern for data

Explainable AI - how is data being used?

Factual AI - response should be backed by data

Flexible in interpreting data

Normative beliefs about AI

- *Models shouldn't make predictions based on demographic characteristics*

- *Model behavior shouldn't be different for different groups of users*

- *Model predictions shouldn't vary based on the person it is making a prediction about*

- *Model performance shouldn't be worse for some groups of users than for others*

- *Models should be able to justify the decisions that they make about people*

Stakeholders

There are different kinds of stakeholders to consider when we talk about the ethics of AI (Bender 2019):

- **Voluntary direct stakeholders**: people who choose to use the system.
- **Involuntary direct stakeholders**: people who must use the system in order to access essential services.
- **Indirect stakeholders**: subjects of queries, contributors to a corpus (voluntarily or involuntarily)
- **Project funders**: the people providing the funding
- **System builders**: the technologists creating the system
- **Communities**: communities impacted by model predictions

The National Science Foundation is considering replacing its peer review system for reviewing grant applications with an automated system. The NSF, together with the NIH, is responsible for funding most of the scientific research conducted at American universities, including directly funding over 100,000 graduate students every year.

graduate students

NSF

NIH

universities

researchers

peer reviewers

communities impacted by research

A farmer is considering adopting a system developed by UC Berkeley computer scientists that uses computer vision to identify pests and zap them with lasers.

pests

crops

farmer

UC Berkeley

computer vision

Computer scientists

consumers

pesticide companies

other farmers

community around farm

EPA

Ag

Roblox, a platform where people can program and share games with each other, is collecting code to train a large language model of code, which they hope will improve the experience of novice programmers. They are using an opt-in mechanism for collecting code.

people who opt in

professional coders

programmers who dont opt in

stock holders

parents of Roblox players

novice programmers

other code platforms:

open source

closed source

Roblox programmers

Roblox as funder

Categorizing Harms

Discussion largely based on Blodgett (2021)

Kinds of Harm

- Allocational harms: *Does the system allocate opportunities or resources unfairly? Do some people gain access more easily than others?*

- Representational harms: *Does this strengthen stereotypes? Does this create or reinforce unfair negative perceptions of a group of people? Does the system fail to even recognize some people?*

Representational Harms

- **Stereotypes**: the system propagates negative generalizations about certain social groups
- **Misrepresentation**: the system performance is skewed towards certain groups of people
- **Erasure**: the system fails to recognize other groups of people
- **Denigration**: the system contains or uses language that is harmful to the dignity or well-being of some people
- **Alienation**: the system denies the relevance of socially meaningful categories

Allocational Harms

- **Quality of service**: the system performs better for individuals who belong to some groups than for others
- **Public participation**: the system makes the speech or contributions of individuals in certain groups less visible than others.
- **Resource allocation**: the system is used in a way that allocates resources more to individuals from one group than another.
- **Opportunity allocation**: the system is used in a way that allocates opportunities more to individuals from one group than another.
- **Targeted surveillance**: the system is used to profile or monitor individuals based on their demographic characteristics.
- **Predictive generalization**: there are disparate impacts across social groups in the treatments/interventions recommended by a system.

Harm Reduction

Microsoft Harms Modeling

Categories of potential harms

- Risk of injury

  physical or emotional

- Denial of consequential services

  opportunity or resource losses

- Infringement on human rights

  losses of privacy, dignity, or liberty

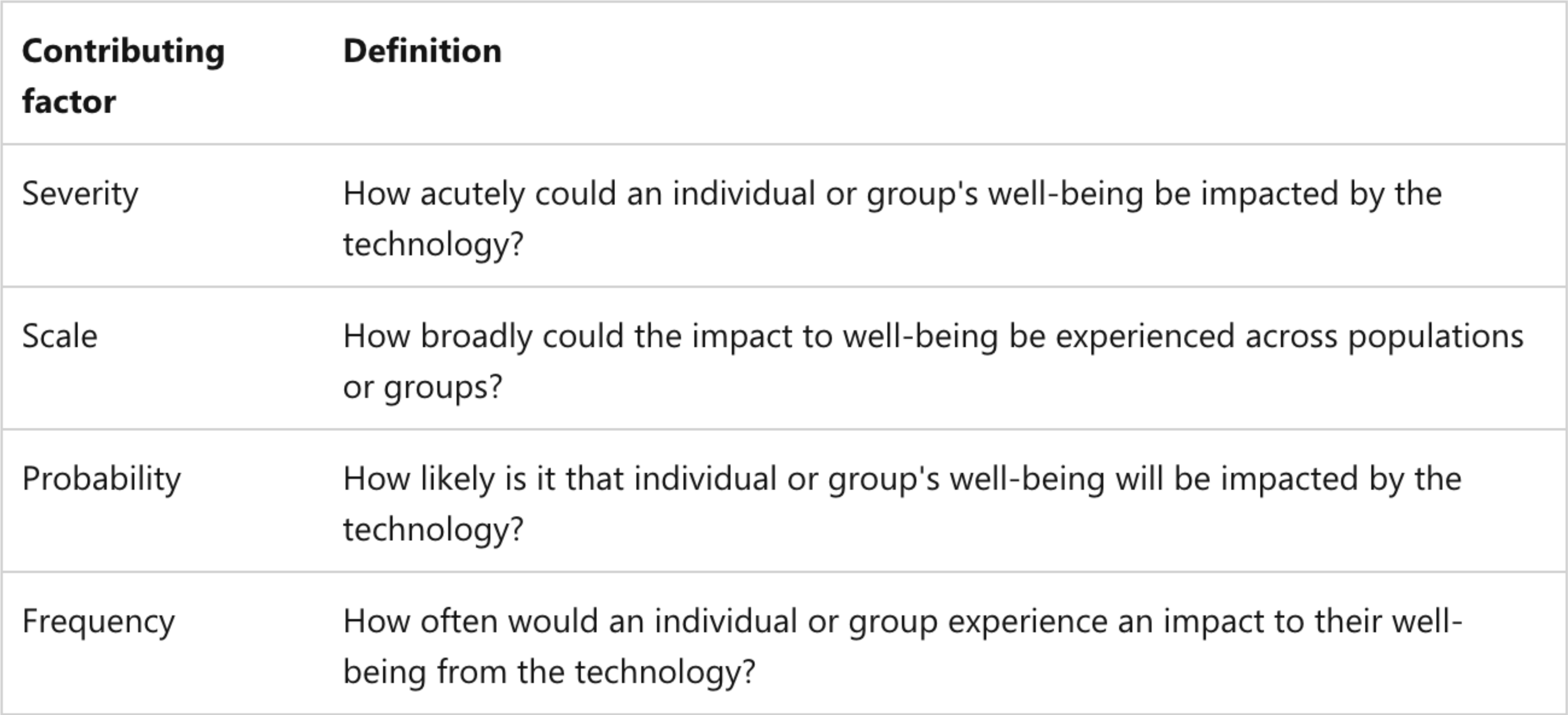
- Environmental impact
- Erosion of social & democratic structures

  misinformation

  social detriment, manipulation, & propaganda

Microsoft Harms Modeling

For each category of harm, consider its:



Contributing factor	Definition
Severity	How acutely could an individual or group's well-being be impacted by the technology?
Scale	How broadly could the impact to well-being be experienced across populations or groups?
Probability	How likely is it that individual or group's well-being will be impacted by the technology?
Frequency	How often would an individual or group experience an impact to their well-being from the technology?

Ethics assessment model: community jury

In the **community jury** model, the potential harms and benefits of a proposed technology are weighed by a diverse group of stakeholders.

- The **product team** creates relevant documentation, data management plan, and prototypes to present.
- A **moderator** facilitates discussion and deliberations.
- A **jury** is assembled of 16-20 community members, sampled in a way that is random but ensures a demographically diverse group.

Ethics assessment model: community jury

2-3 hr sessions are held to assess the proposed technology:

- **Overview and introduction**: The moderator explains the rules of conduct. The product team explains the proposed technology and its goals.
- **Q&A**: jurors ask questions about the technology.
- **Deliberation and cocreation**: the jury and product work together to come up with solutions that meet all needs.
- **Anonymous surveying (optional)**: anonymously poll the jurors to get their honest opinions.
- **Study report**: the moderator writes a report outlining key insights, concerns, and proposed solutions.

Scenario: Code Generation

Roblox, a platform where people can program and share games, is collecting code to train a large language model of code. Their goal is to improve the experience of novice programmers.