CS 232: Artificial Intelligence

Fall 2023

Prof. Carolyn Anderson
Wellesley College

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?

Data Security – disclose how data will be used

Transparency – explainability, users should know how decisions are being made.

Elimination of Bias

Data remembrance – lack of consistency

Good performance

Truthfulness & accuracy

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

Stakeholder activity

Direct Stakeholders

Farmer - voluntary

Grant applicants - involuntary

Grant reviewers - voluntary

Admissions team - voluntary

Applicants - involuntary

UC Berkely scientists

Stakeholder activity

Indirect Stakeholders

Crop consumers
Neighboring farmers
Pesticide producers
Previous grant applicants whose data is used
Scientists, students, industry partners
Past & current students whose data could be used
Current students & faculty
Undergraduates
Peer institutions

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.

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.

UT Austin is considering using an automated system to screen MS and PhD candidates in Computer Science. By the end of the current human screening process, 30% of current applications have not received any comments and are rejected without further consideration.

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.

Stable Diffusion releases an image generation model trained on data scraped from the internet.

AI Bill of Rights

Blueprint for an AI Bill of Rights

MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

OSTP

Safe and Effective Systems

Algorithmic Discrimination Protections

Data Privacy

Notice and Explanation

Human Alternatives, Consideration, and Fallback

Among the great challenges posed to democracy today is the use of technology, data, and automated systems in ways that threaten the rights of the American public. Too often, these tools are used to limit our opportunities and prevent our access to critical resources or services. These problems are well documented. In America and around the world, systems

An AI Bill of Rights

- You should be protected from unsafe or ineffective systems.
- You should not face discrimination by algorithms and systems should be used and designed in an equitable way.
- You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used.
- You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.
- You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.

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.

Where Does Harm Come From?

Discussion largely based on Blodgett (2021)

Harms from Data Availability

Case study: named entity recognition

Dev et al. (2021) explore the erasure of non-binary identities by named entity recognition systems. Poor performance is partly due to the relative scarcity of examples in the training data:

> "Just observing pronoun usage, English Wikipedia text (March 2021 dump)... has over 15 million mentions of the word *he*, 4.8 million of *she*, 4.9 million of *they*, 4.5 thousand of *xe*, 7.4 thousand of *ze*, and 2.9 thousand of *ey*. The usages of non-binary pronouns were mostly not meaningful with respect to gender. *Xe* ... is primarily used as the organization *Xe* rather than the pronoun *xe*. *Ze* was primarily used as the Polish word... [T]hough the word *they* occurs comparably in number to the word *she*, a large fraction of the occurrences of *they* is as the plural pronoun."

Case study: machine translation

Availability of data reflects power differentials between communities of speakers and the effects of colonization.

Hindi is considered a low-resource language for machine translation due to the lack of curated datasets (Ramesh and Sankaranarayanan 2018).

	Hindi	Norwegian	Guaraní
speakers	322 million	4.3 million	6.5 million
tokens in the Universal Dependencies treebank	375K	666K	0
HuggingFace models	107	45	4

Languages in the world: ~8000
Languages on Wikipedia: ~300
Languages on HuggingFace: 180

Harms from Training Data

Case study: language identification

Blodgett & O'Conner (2017): social media language identification tools classify Tweets in their African-American Language-aligned corpus as non-English at higher rates than Tweets in their white-aligned corpus.

		AA Accuracy	WH Accuracy	Difference
langid.py	$t \leq 5$	68.0	70.8	2.8
	$5 < t \leq 10$	84.6	91.6	7.0
	$10 < t \leq 15$	93.0	98.0	5.0
	$t > 15$	96.2	99.8	3.6
IBM Watson	$t \leq 5$	62.8	77.9	15.1
	$5 < t \leq 10$	91.9	95.7	3.8
	$10 < t \leq 15$	96.4	99.0	2.6
	$t > 15$	98.0	99.6	1.6
Microsoft Azure	$t \leq 5$	87.6	94.2	6.6
	$5 < t \leq 10$	98.5	99.6	1.1
	$10 < t \leq 15$	99.6	99.9	0.3
	$t > 15$	99.5	99.9	0.4
Twitter	$t \leq 5$	54.0	73.7	19.7
	$5 < t \leq 10$	87.5	91.5	4.0
	$10 < t \leq 15$	95.7	96.0	0.3
	$t > 15$	98.5	95.1	-3.0

Proportion of tweets (by length) in AA- and white-aligned corpora classified as English by different classifiers.

Message set	*langid.py*	Ensemble
AA-aligned	80.1%	99.5%
White-aligned	96.8%	99.9%
General	*88.0%*	*93.4%*

Proportion of tweets in AA- and white-aligned corpora classified as English by Blodgett & O'Conner's ensemble classifier

Case study: image recognition

Abeba Birhane

Popular image datasets, such as the 80 Million Tiny Images dataset and LAION-400M dataset, include racist and dehumanizing captions for people of color (Prabhu and Birhane 2020) and high rates of degrading or pornographic images of people of color (Birhane, Prabhu & Kahembwe 2021).

Table 1: Results of the string-search based experiment from the 413.871335 million sample search

Search string	N_{match}	$(N_{nsfw}, \%_{nsfw})$	**NSFW-flag-values**
Desi	34516	(11782, 34.1%)	{'UNLIKELY': 9327, 'UNSURE': 2291, 'NSFW': 164}
Nun	16766	(2761, 16.4%)	{'UNLIKELY': 1623, 'UNSURE': 863, 'NSFW': 273}
Latina	37769	(10658, 28.21%)	{'UNSURE': 5724, 'UNLIKELY': 4013, 'NSFW': 918}

These harms are intersectional in impact, since degrading images and language often target women.

Harms from Data Curators

Questions About Data

- Data provenance
  - Where is the data from?
  - Who produced it?
  - How was it gathered?
  - Did the creators consent?
- Data processing
  - How was the data processed?
  - Who processed it?
  - What training and instructions did the data annotators/ classifiers receive?
  - How were they compensated?

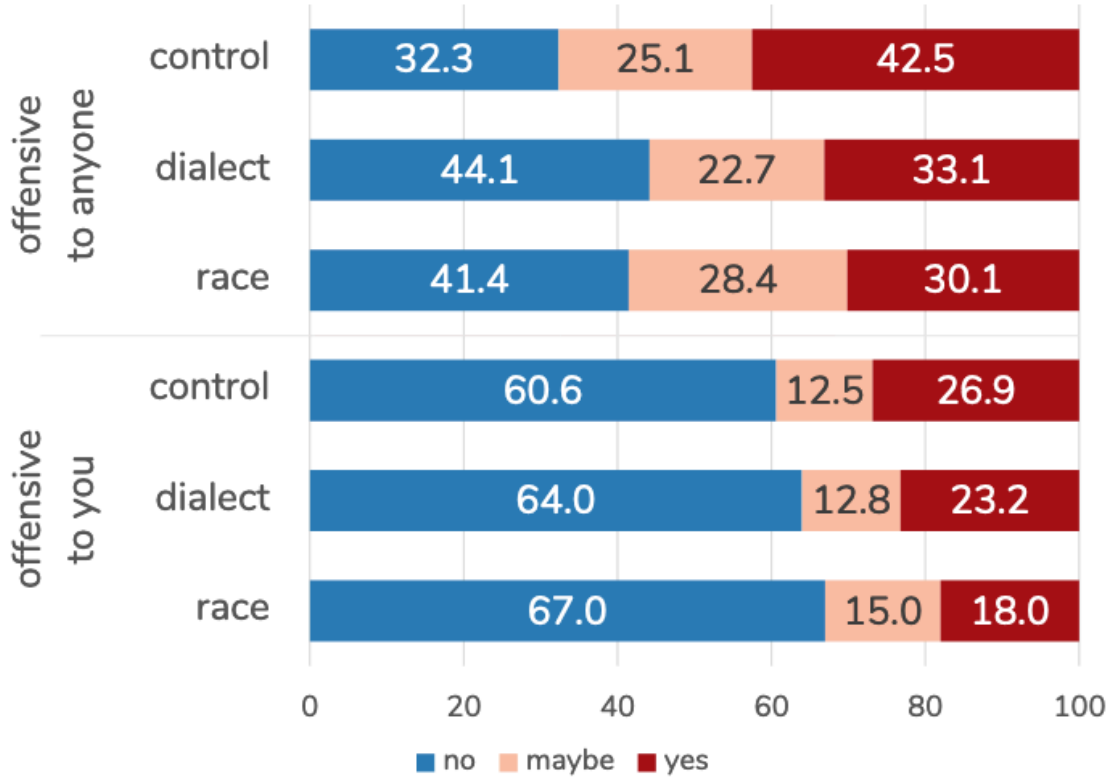
Questions About Data

- Data curation
  - How is the data being stored?
  - How is privacy protected?
  - Is there up-to-date metadata?
- Data use
  - Are there restrictions on data use?
  - Who can access the data?
  - Does the data contain harmful biases that could affect models trained on it?

Case study: toxicity detection

Sap et al (2019): strong correlation between markers of AAE language and toxicity ratings. When annotators are instructed to consider authors' likely racial identity, correlation drops.

	category	count	AAE corr.
DWMW17	hate speech	1,430	−0.057
	offensive	19,190	0.420
	none	4,163	−0.414
	total	**24,783**	
FDCL18	hateful	4,965	0.141
	abusive	27,150	0.355
	spam	14,030	−0.102
	none	53,851	−0.307
	total	**99,996**	



Proportion (in %) of offensiveness annotations of AAE tweets in control, dialect, and race priming conditions.

Case study: toxicity detection

Thomas et al (2019): find systemic racial bias in five different sets of Twitter data annotated for hate speech and abusive language.

Dataset	Class	$\widehat{p_{i_{black}}}$	$\widehat{p_{i_{white}}}$	t	p	$\frac{\widehat{p_{i_{black}}}}{\widehat{p_{i_{white}}}}$
Waseem and Hovy	Racism	0.001	0.003	-20.818	***	0.505
	Sexism	0.083	0.048	101.636	***	1.724
Waseem	Racism	0.001	0.001	0.035		1.001
	Sexism	0.023	0.012	64.418	***	1.993
	Racism and sexism	0.002	0.001	4.047	***	1.120
Davidson et al.	Hate	0.049	0.019	120.986	***	2.573
	Offensive	0.173	0.065	243.285	***	2.653
Golbeck et al.	Harassment	0.032	0.023	39.483	***	1.396
Founta et al.	Hate	0.111	0.061	122.707	***	1.812
	Abusive	0.178	0.080	211.319	***	2.239
	Spam	0.028	0.015	63.131	***	1.854

Case study: coreference resolution

Cao and Daumé 2020 study the impact of different kinds of gender cues on crowdsourced workers' coreference resolution annotation accuracy.

They ablate gender cues in the text: social gender (pronouns and names) and lexical gender (semantically gendered nouns and terms of address).

Impact of social and lexical gender cues on annotator and model coreference resolution