Ethics in Data Science?

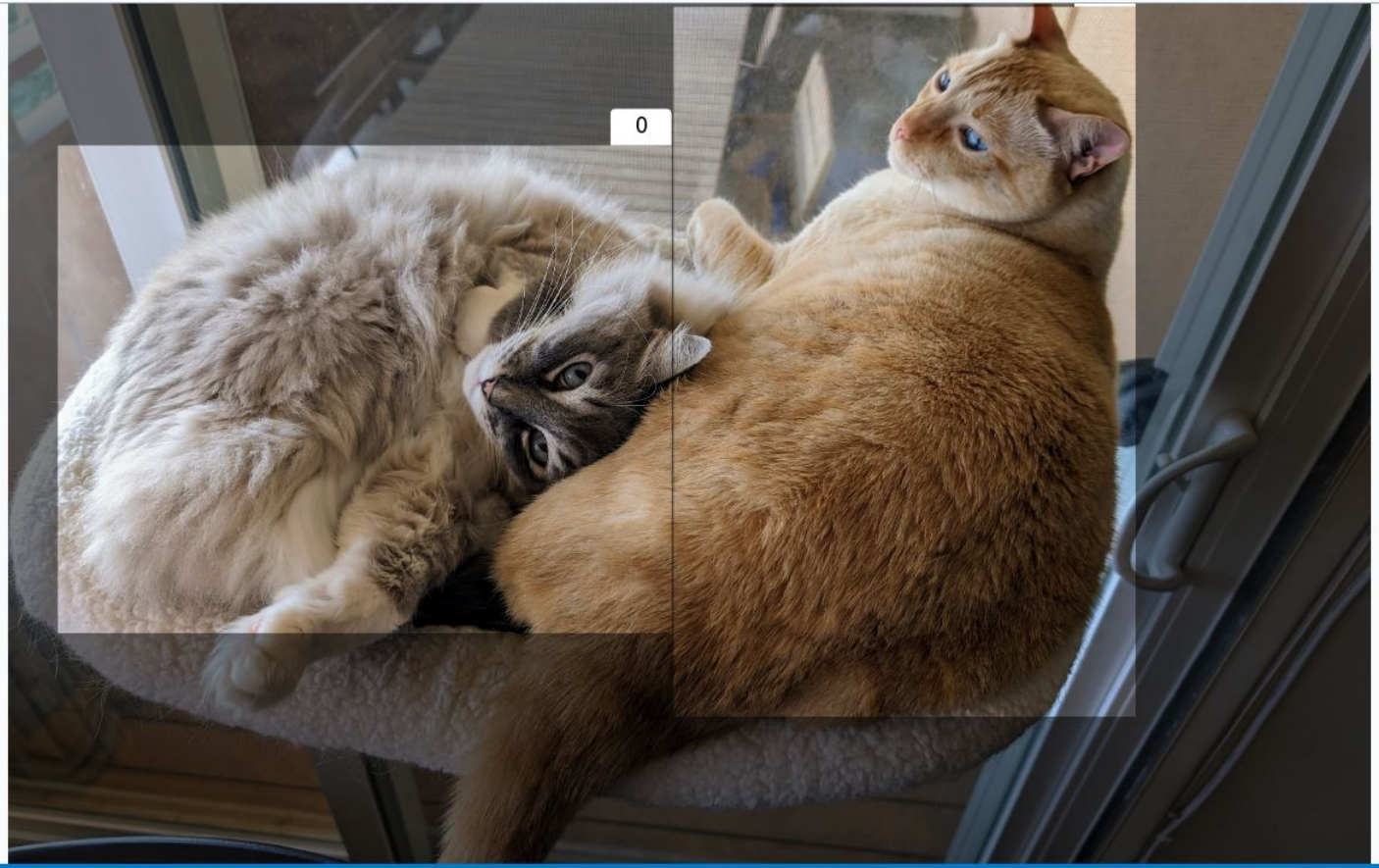
DATA 120 (Spring 2026)

Justin Sola, Assistant Professor of Sociology and
School of Data Science and Society

How are you feeling?

0

Which cat is cuter?



Form a group of up to 4 people

1. Introduce yourselves
2. Determine which cookout shake flavor is best
3. Declare it in the following PollEverywhere

Group exercise: what is your group's favorite cookout milkshake flavor?

Nobody has responded yet.

Hang tight! Responses are coming in.

Today

- What is ethics?
- Why is data science a special domain for ethics
- A tour of where ethics shows up across the DS lifecycle

Anchor idea: data systems are not just code – they are socio-technical and redistribute power, risk, and attention.

Part 1 – Definitions

How can we start thinking about ethics?

1. **Ethics**: disciplined, reason-giving reflection on what we *ought* to do—and why—especially when **values collide**.
2. Why does this matter in DS/AI?
3. Three layers:
   a. **Normative** — What actions/policies are right or wrong?
   b. **Meta-ethics** — What do "right," "good," or "fair" even mean?
   c. **Applied** — Consider values in tension in real domains (e.g., data science), and be explicit about practical trade-offs.

Ethics vs. Morals

"We are discussing no small matter, but how we ought to live" – Socrates

Morals

- Personal norms and intuitions about right/wrong.
- Shaped by upbringing, community, culture.
- Often tacit (“this just feels wrong”).

Ethics

- Collective agreement around those norms.
- Shaped by collective reasoning, debate, and shared values.
- Can be explicit, but is mostly understood when considered in light of the communal well-being.
  - Offers frameworks for reasoning about moral questions!

Ethics

- Collective agreement around those norms.
- Shaped by collective reasoning, debate, and shared values.
- Can be explicit, but is mostly understood when considered in light of the communal well-being.
  - Offers frameworks for reasoning about moral questions!

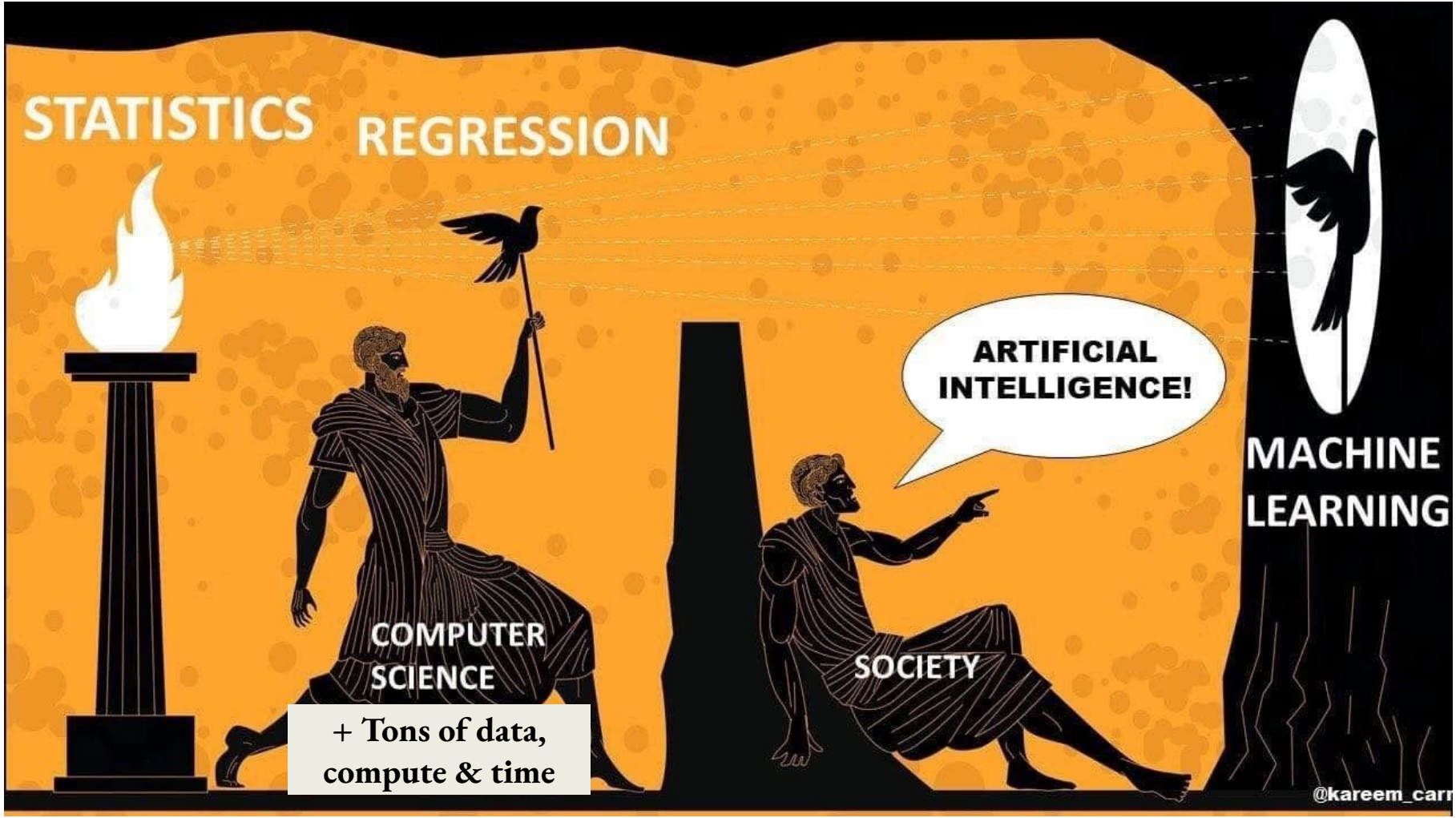
Morals

- Personal norms and intuitions about right/wrong.
- Shaped by upbringing, community, culture.
- Often tacit (“this just feels wrong”).

What is data science? What is artificial intelligence?

Data science vs. ‘AI’

- Data science (data collection, **analysis**, and insight generation)
- ‘Artificial Intelligence’ (a wide set of new analytic techniques)
  - Artificial intelligence is often too superficial a term
  - A random forest…
  - …is different from a computer vision classifier
  - …is different from a generative large language model


STATISTICS
REGRESSION
ARTIFICIAL INTELLIGENCE!
MACHINE LEARNING
COMPUTER SCIENCE
SOCIETY
+ Tons of data, compute & time
@kareem_carr

What is the application of these methods?

Many ML methods—though they can be far more elaborate than simple linear regression—share the same foundational concept of regression:

fitting functions to data to make ***predictions or inferences***

Key Areas of Impact of these inferences:

- Healthcare: AI-assisted diagnostics, patient data analytics
- Criminal Justice: Predictive policing, sentencing algorithms
- Finance: Credit scoring, fraud detection
- Social Media: Recommendation systems, content moderation

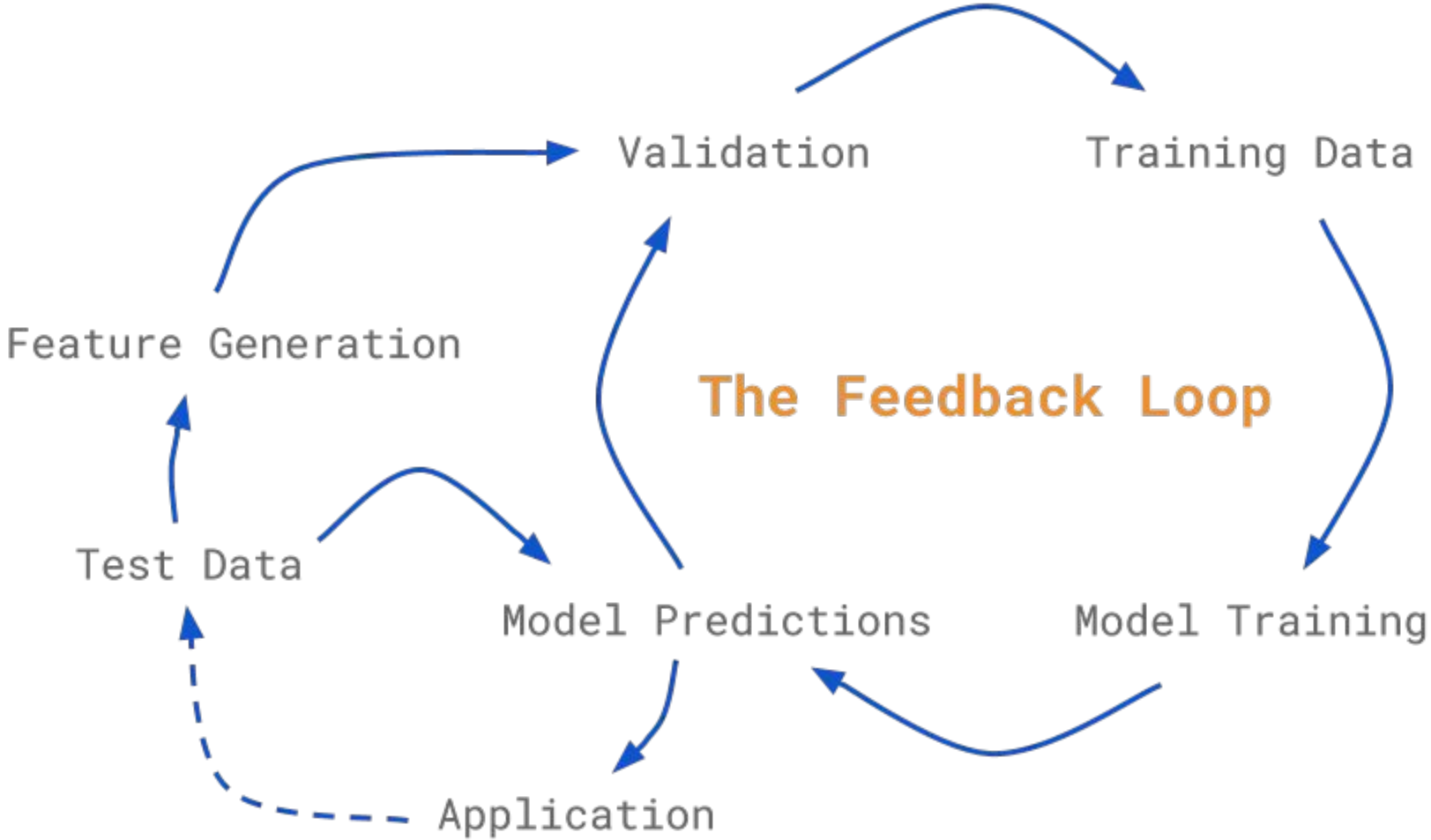
Why Ethics Matters in Data Science

Ethical Dilemmas:

- Privacy concerns, data biases, and the misuses of these methods
- Impact of algorithmic decisions

Core Questions:

- What is fairness in AI systems?
- How do we identify and mitigate societal harms?
- Reference real-life controversies


Validation
Training Data
Feature Generation
The Feedback Loop
Test Data
Model Predictions
Model Training
Application

Case Study: Algorithmic Biases from Endogeneity

A feedback loop in a data generation process may be problematic when decisions made on the basis of data in turn *affect future data collection*. Example:

1. You note that crime is geographically concentrated
2. You recommend more police presence in such areas, and police redistribute
    a. Police make fewer arrests where they are not, and more arrests where they are
3. Arrests are *endogenous* (influenced by factors within the system being analyzed) rather than *exogenously* related to crime
4. In comparison, 911 calls and victimization surveys are not (*as*) directly affected

Cruise admits to filing false report after robotaxi dragged a San Francisco pedestrian

By **Andrew Mendez** • Published November 14, 2024 • Updated on November 14, 2024 at 7:17 pm

Part 2 – Introductory Applications

What Do You Think Fairness Means? Could be…

- equal distribution (everyone gets the same resources),
- equal opportunity,
- first-come-first-served (queuing),
- demographic parity / representation,
- effort-based (‘you get out what you put in’),
- market distribution,
- progressive distribution (e.g., children and elderly get a leg up),
- merit distribution,
- combinations thereof, etc.

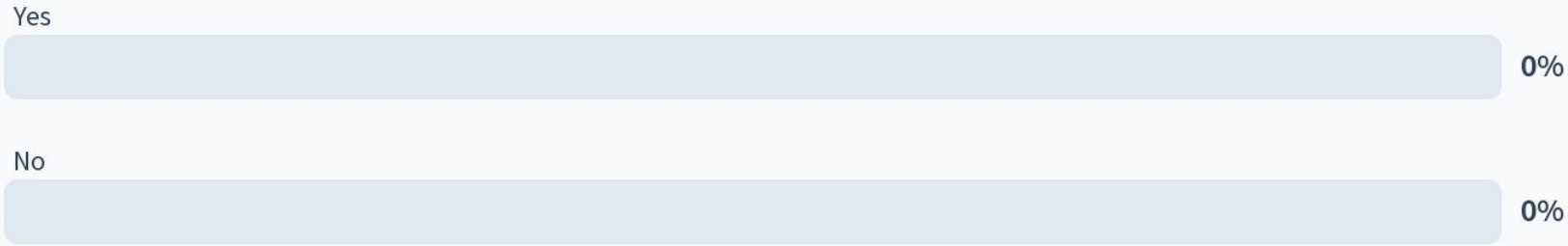
Brief group exercise (3 minutes)

Imagine the following situations:

1. Coffee Shop: A person asks to skip the line because they only need hot water.
2. Emergency Room: A patient with mild symptoms arrives first, but a second patient with life-threatening symptoms arrives later.

Should we use the same algorithm to allocate resources for both situations? Discuss among your group for 1 minute.

Same algorithm for coffee shop and emergency room allocation?



Group exercise 1 (~4 minutes)

1. Pick 1 of 3: grades in a course, medical care at an ER, drinks at a coffee shop
2. Discuss *the simplest fair way* to distribute the resource or metric in question
    a. Note disagreements (those are good!)
3. Post that simple (1 sentence) metric and upvote/downvote peer answers in this format: “team_name course/coffee/ER: fairness metric”

Fairness metric group response

Nobody has responded yet.

Hang tight! Responses are coming in.

Part 3

What’s the point?

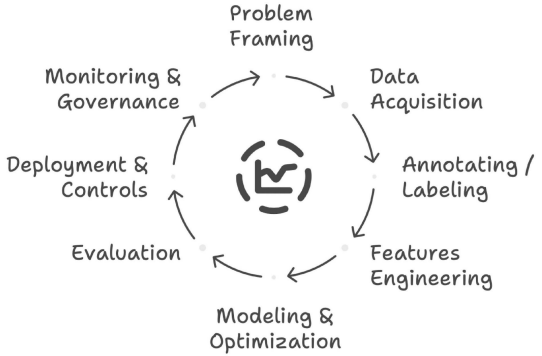
- There usually isn’t a clearly superior answer to a moral dilemma
  - If there were, it wouldn’t be much of a dilemma!
- Our values differ, our resources are constrained, and even our capacity to articulate our preferences is limited
  - Worthwhile to discuss and highlight such dilemmas

Daily Dilemma

If there was no difference in the accuracy of the outcome, would it matter to you if you were judged by a machine or by a human? Why or why not? Specifically, which concepts/values activate for you when forming and answer?

The DS lifecycle (and where ethics lives)

Stages
- Problem framing
- Data acquisition
- Annotating / labeling: ground truth?
- Features engineering / representation
- Modeling & objectives being optimized
- Evaluation
- Deployment & controls
- Monitoring & governance



Values to keep in mind

- Autonomy & freedom
- **Justice & fairness**
- **Transparency & explanation**
- Beneficence & non-maleficence
- Responsibility & accountability
- **Privacy**
- Trust
- Sustainability, dignity, solidarity

These interlock; cases rarely isolate just one!

- Questions to consider:
  - What goals are being pursued? By whom?
  - What does success achieve? What cost does it impose? Are they proportional?
  - What cannot be measured here, and what's the cost of pretending we can?
  - Who is absent from the room (stakeholders)?
- **Which values are in tension?**

Data collection & consent

- Surveillance that “helps” can still impose costs and change behavior.
- Trade-offs: safety vs. dignity; consent vs. practicality; individual vs. caregivers/institutions.
- Always ask: *Who watches? Who benefits? Who bears risk?* Case: monitoring people living with dementia—useful, but privacy, agency, and dignity are relevant issues.

Labeling, ground truth, & proxies

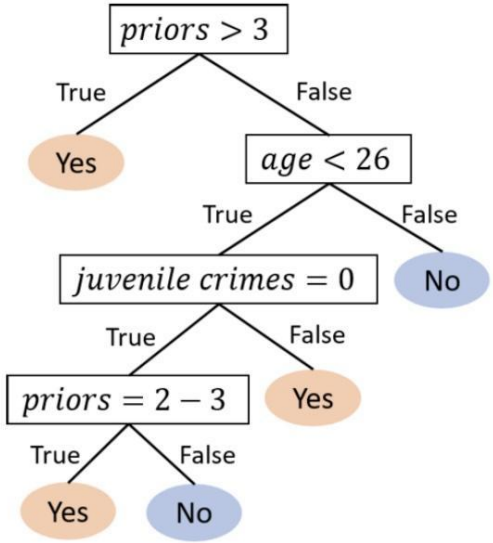
- Many labels (e.g., "risk," "toxicity," "creditworthiness") are social judgments.
- Ground truth $\neq$ objective truth; be curious about how labels are produced.
  - Example: "toxicity" in language models is often a social judgment.
- Careful! suspicion can be corrosive, but so is blind trust.
  - "Everything is ideology" is as sloppy as "the math is objective."
- - Sensitive attributes often sneak back in via proxies (ZIP code, devices, networks).

Modeling objectives

- Loss functions encode values: false negatives vs. false positives.
- Class weights, thresholds, regularizers — who is protected vs. exposed?
- Be explicit: why *this* objective for *this* domain?
- Consider ways in which evaluation can move beyond accuracy

Transparency: interpretation and explanation

- Interpretability != explainability...
- "Right to an explanation" vs. model complexity (*not* accuracy!)
- Explanation should fit the audience (auditor vs. end user).
- Over-transparency invites gaming; under-transparency erodes trust.

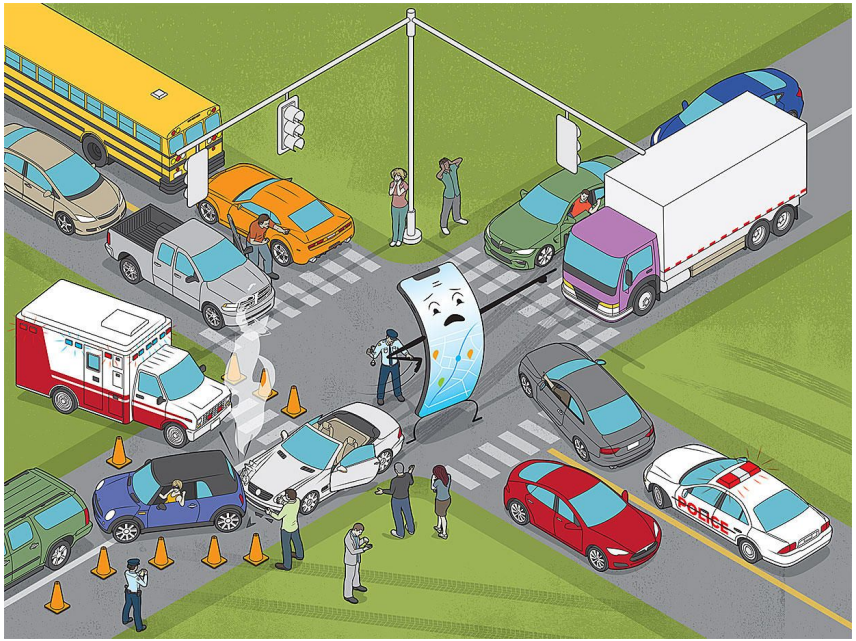


(a)

Deployment & “human-in-the-loop” limits

- Controls depend on context and reaction time.
- In some domains, a human can't intervene fast enough; in others, post-hoc review is fine.
- Opacity & complexity can blunt human oversight.
  - Build guardrails appropriate to the risk.

Feedback loops & drift (socio-technical)



- Data → model → behavior → new data
- Political messaging, recommender systems, policing, credit — all exhibit looped reinforcement
- Governance must anticipate how using a model changes the world it measures.

Part 4

How to read academic sources

Dilemma

- You need to do the readings…
- BUT they are dense!

I strongly suggest you read Rubin 2019

Particularly the notes template on the last 2 pages!!! Easy mode

How to read – pt 1

Author's Tone

- ☐ Empirical/Neutral
- ☐ Normative/Biased
- ☐ Policy Recommendation
- ☐ Other _____

Data and Method – Check All that Apply

- ☐ Ethnographic/Participant Observation
- ☐ Interview
- ☐ Content Analysis
- ☐ Historical/Archival Records
- ☐ Historical Narrative
- ☐ Unclear
- ☐ Survey Data
- ☐ Census Data
- ☐ Regression or Other High-Level Statistical, Computer-based Analysis
- ☐ Trends, Averages, Counts, or Basic Descriptive Statistics
- ☐ Other _____

How to read – pt 2

What are the…

1. research question(s) of the study, and expectations about them
2. research method(s) of the study
3. finding(s) of the study, and how the author(s) interpret these findings
4. Links between the reading and your other sources of knowledge