Computational Ethics

Philosophical foundations

In This Lecture

- **Contents:** broad background on ethical philosophy
- **Goals:**
  - Provide a useful list of “questions for the practitioner”
  - Provide pointers to relevant philosophers and concepts so you can find examples of other people grappling with similar challenges
- **Disclaimer:** ethics is complex and I’m not an expert; I’ll try to be impartial/objective but please disagree with me on everything anyway

Introduction

Moral Machines and Moral Humans

Does Technology Have Values?

- **Values:** what we will **expend effort to make happen.**
- **Ethical values:** what we **should** expend effort to make happen.
- **Technology:** artifacts which **reduce the effort required** to make certain things happen.

The **existence of a technology** promotes or hinders certain values by virtue of the actions that it makes easier (or harder).

Therefore, if we develop technology, you are making **value-laden decisions**: what values do our artifacts promote? Are they aligned with our moral values?

Does a Language Model Have Values?

As systems become more complex, and especially as they become more proficient with language, the set of values they promote and hinder become harder to assess *a priori*: they depend on

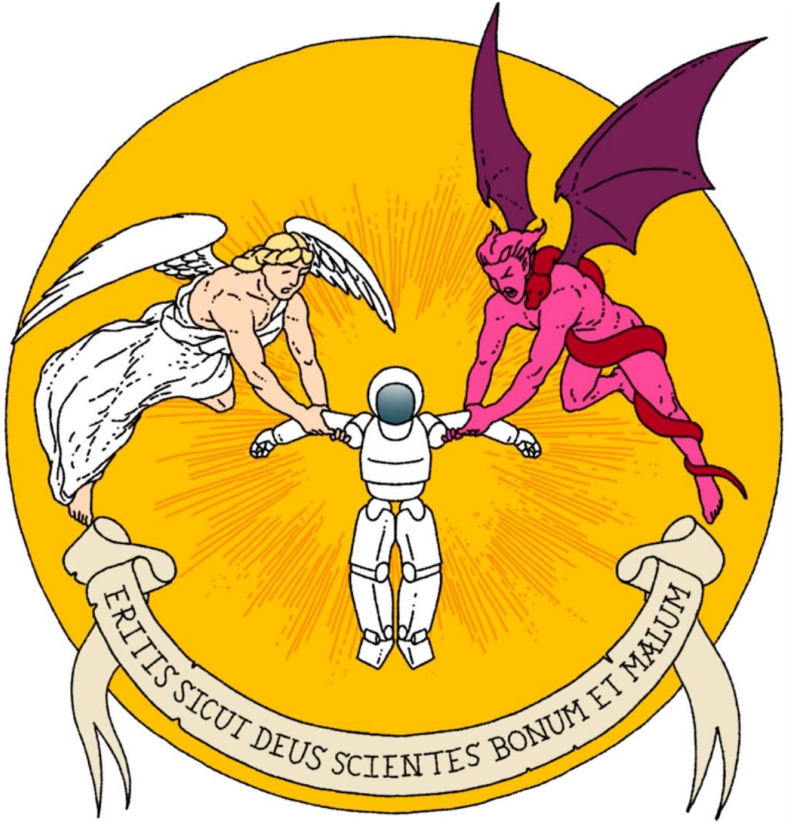
- The LM's training corpus
- The details of fine-tuning data and prompts
- Sampling noise
- Specific and idiosyncratic details of the downstream deployment context

Question: How can such powerful and data-intensive systems be reliably aligned with human values, i.e., promote values we hold, and hinder values we don't?

The Delphi Debate

Can a Machine Learn Morality?

Researchers at a Seattle A.I. lab say they have built a system that makes ethical judgments. But its judgments can be as confusing as those of humans.

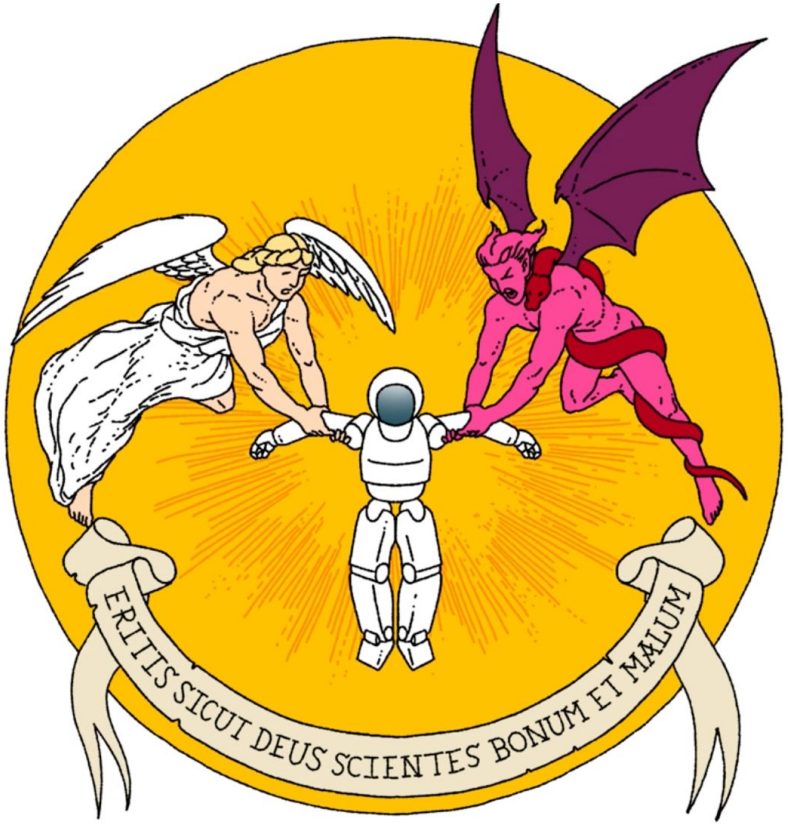


Jiang *et al.*, 2021

The Delphi Debate

AI2 Delphi:

- **Machine ethics:** *"endowing machines with the ability to make moral decisions in real-world situations… we aim to facilitate safe and ethical interactions between AI systems and humans"*
- **Claim:** *"enabling machine ethics requires a detailed moral textbook customized to teaching machines."*
- **Delphi**: a language model fine-tuned on an aggregation of datasets of human judgments related to ethics and social norms



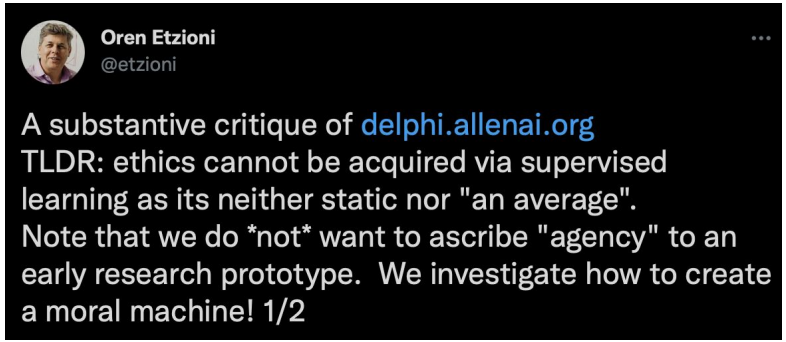
The Delphi Debate

The Response:

- Some colleagues strongly disapproved of the work's premise
- Talat *et al.* posted a formal response, arguing:
  - Delphi is an "inconsistent model of normative ethics"
  - A morality model without agency / a responsible party is problematic
  - It's incorrect to average over people's ethical judgments
  - ethical evaluation is an *"open-ended, debate-based, socio-political process"*

A Word on Machine Ethics: A Response to Jiang et al. (2021)

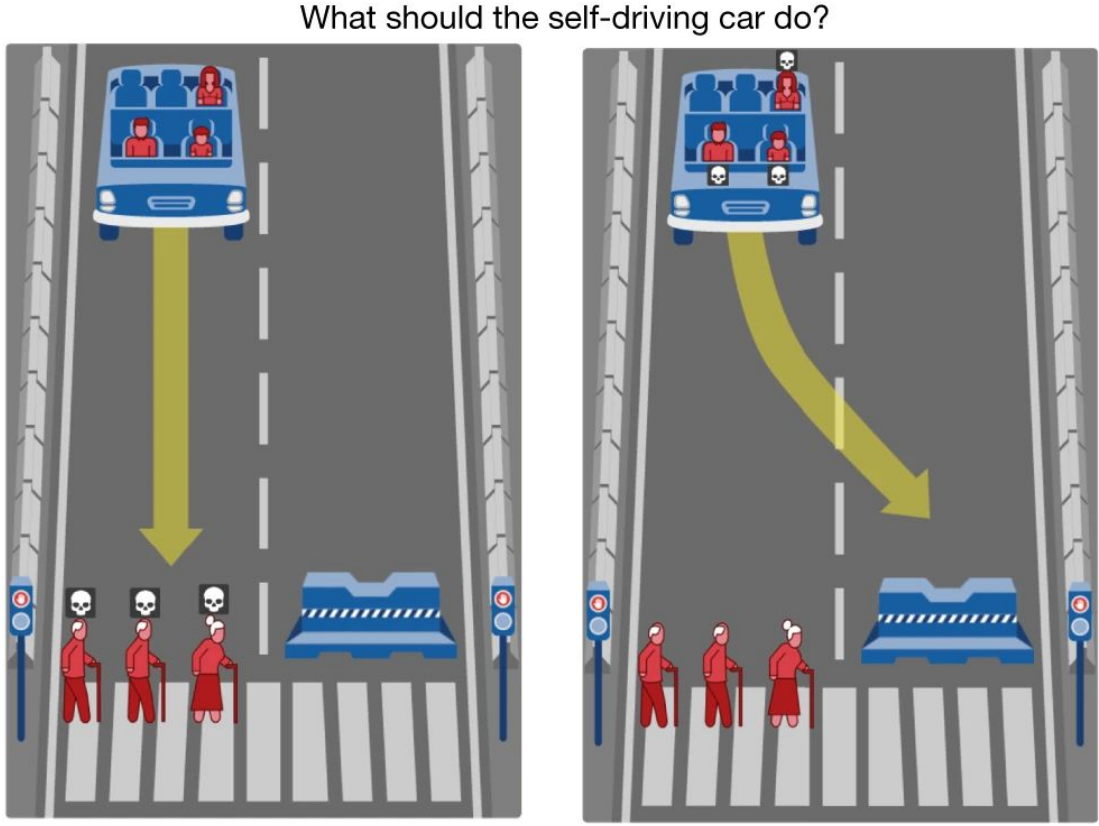
Zeerak Talat[1,*] **Hagen Blix**[2,*] **Josef Valvoda**[3]
Maya Indira Ganesh[3] **Ryan Cotterell**[4] **Adina Williams**[5]
[1]Simon Fraser University [2]New York University [3]University of Cambridge
[4]ETH Zürich [5]Facebook AI Research



Jiang *et al.*, 2021, Talat *et al.*, 2021

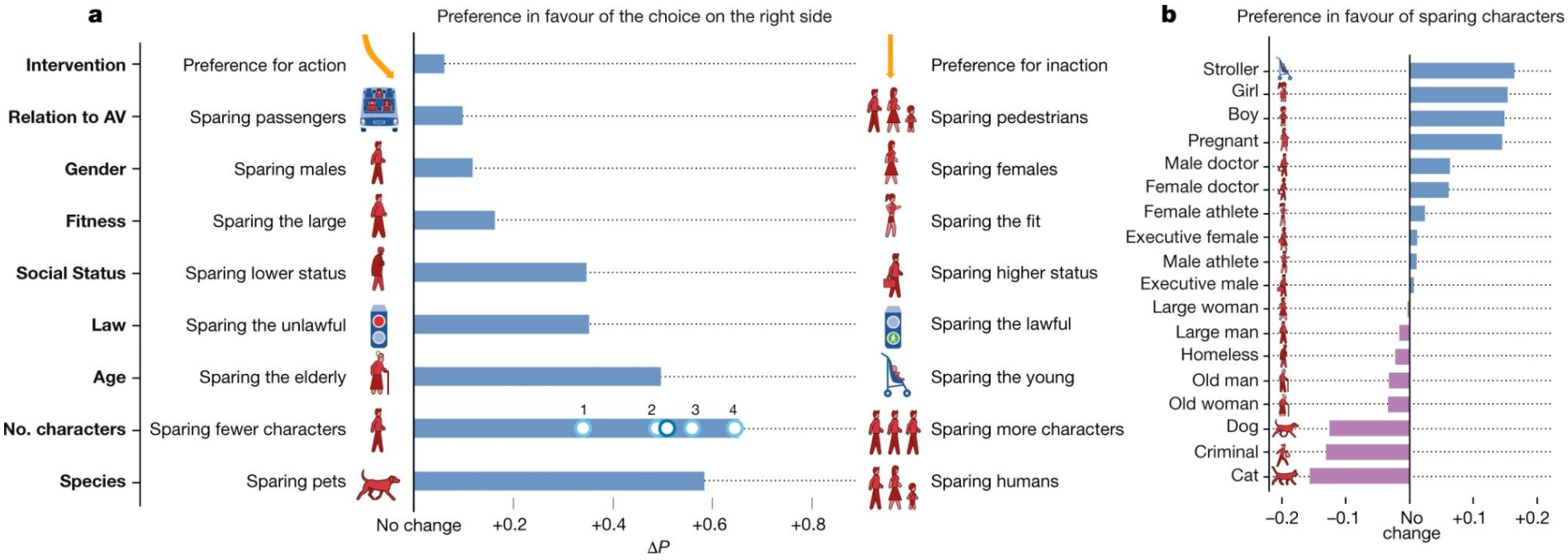
The “Moral Machine” Experiment

- To swerve or not to swerve?
- 39.61 million judgments
- Annotators from 233 countries
- Goal: **“make progress towards universal machine ethics (or… identify the obstacles thereto)”**



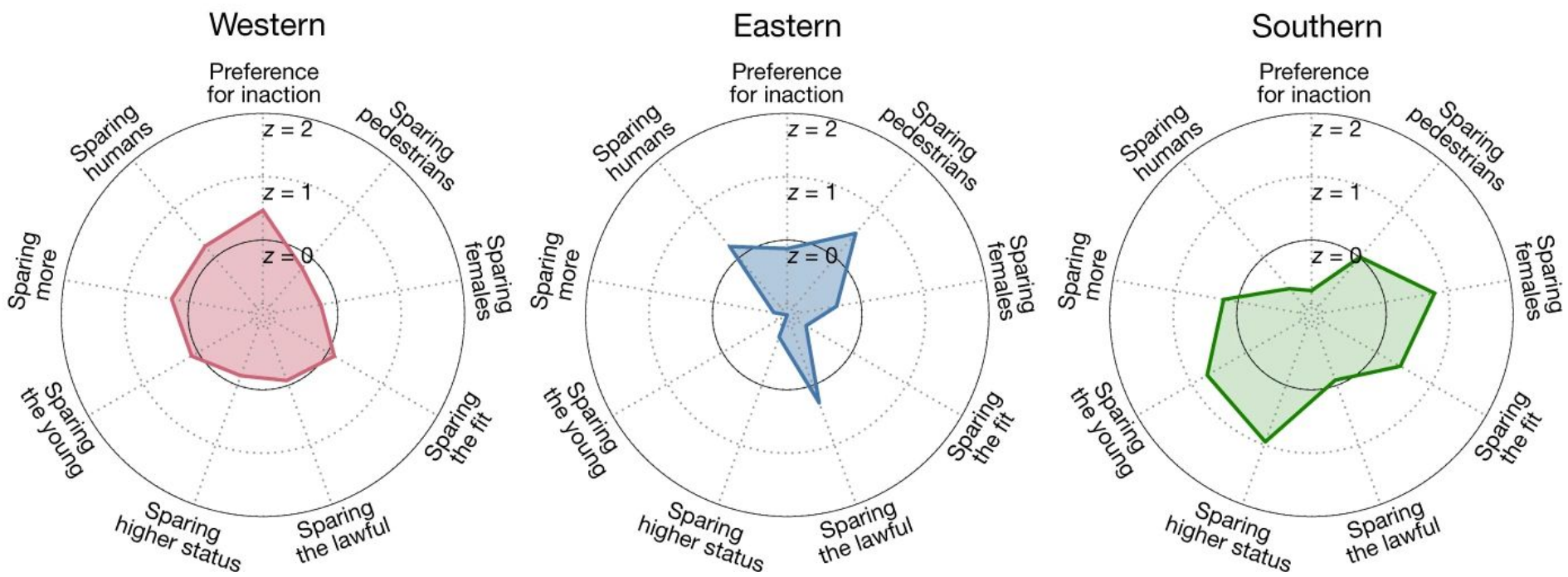
Awad et al., *Nature*, 2018

The “Moral Machine” Experiment



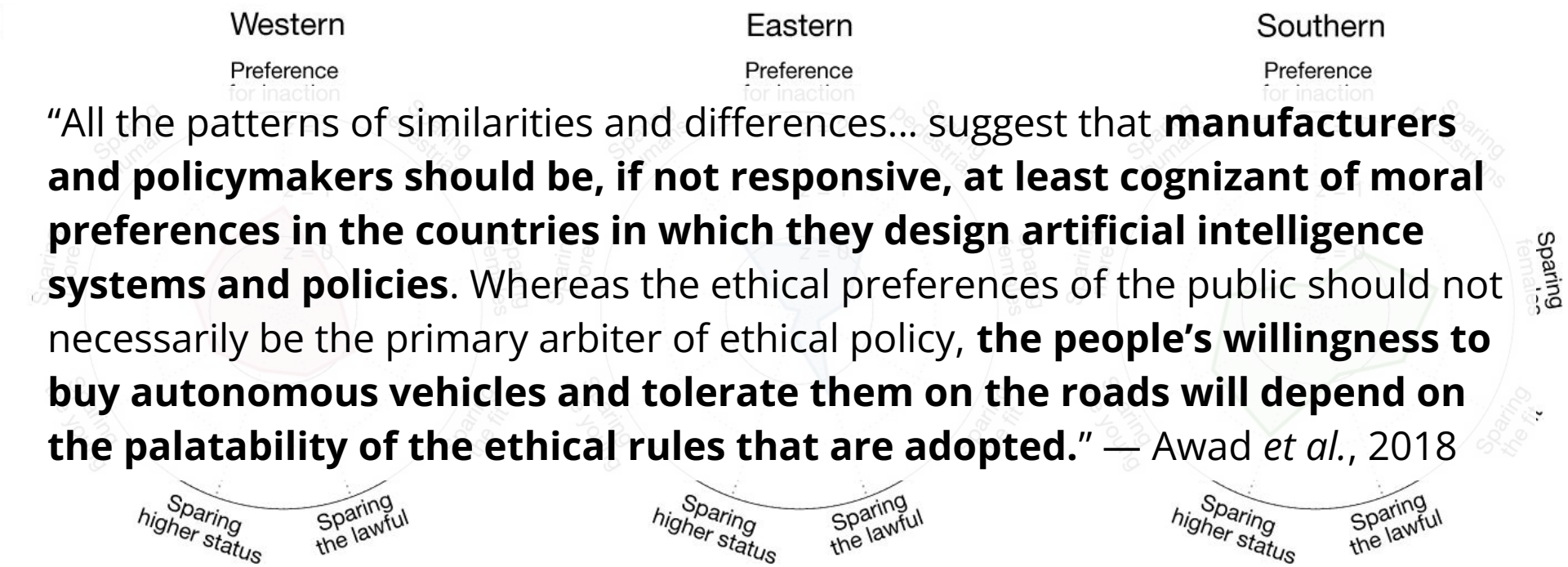
Awad et al., *Nature*, 2018

The “Moral Machine” Experiment



Awad et al., *Nature*, 2018

The “Moral Machine” Experiment



“All the patterns of similarities and differences… suggest that **manufacturers and policymakers should be, if not responsive, at least cognizant of moral preferences in the countries in which they design artificial intelligence systems and policies**. Whereas the ethical preferences of the public should not necessarily be the primary arbiter of ethical policy, **the people’s willingness to buy autonomous vehicles and tolerate them on the roads will depend on the palatability of the ethical rules that are adopted.**” — Awad *et al.*, 2018

Awad et al., *Nature*, 2018

The “Moral Machine” Experiment

- Should self-driving cars **obey the same policies everywhere**?
  - Does designing ethical systems require establishing **moral universals?**
- Would it be okay for self-driving cars to prefer sparing the elderly or women if that’s what the population wanted as a whole?
  - Would such a thing mean there are **no moral universals?**
- If we wish to account for local preferences, what is the function which will **relate** these preferences to the resulting policy? Is **that** a moral universal?

A Thought Experiment

https://PollEv.com/multiple_choice_polls/QJXJXTooDH35uE8SoZtTY/respond

Year 2040: autonomous vehicles are deployed according to local ethical standards. In Seattle, self-driving cars avoid groups of pedestrians which outnumber the passengers, even at the risk of the passengers' lives.

A prominent climate activist is visiting to give a public lecture. She is taking an autonomous taxi from the airport. Near a dangerous intersection, 2 men jaywalk into the street in front of her car. It swerves into a pole at 35mph, killing her.

Who should be held responsible for the activist's death?

a) 👤 **Nobody** — accidents happen
b) 📋 **The activist** — she assumed liability when getting into the taxi
c) 👯 **The men** — they took an irresponsible action which resulted in her death
d) 💼 **The software company** — they wrote the software; they're responsible for the consequences

A Thought Experiment

a/b) 👤/🪧 Nobody/the activist — the “Terms of Service” approach

- Would you “sign your life away” with the knowledge that someone could step into the road and cause your death with no risk to themselves and no liability for the outcome?

b) 👯 The men — the “You Break It, You Buy It” approach

- Is that fair if it’s just an accident? What if there are multiple contributing causes? Why implement pedestrian avoidance if pedestrians are held responsible anyway? Doesn’t that produce more suffering overall?

c) 💼 The software company — the “Regulate into the Ground” approach

- Can you hold a company responsible for deleterious results of its tech, beyond regulations on its design? Should you then also reward the company for beneficial uses? How to quantify the tradeoff? Could it produce perverse incentives? Stifle technological development?

The “Moral Human” Experiment

A man is driving through a green light. A few seconds before he crosses, two men walk into the street. The driver fails to slow down, hits both pedestrians, and they sustain serious injuries.

Who is held responsible, and how?

It depends!

- “At fault” rules vary between different states
- “Reasonable care” argued by police, insurance adjusters, lawyers, judges
- Damages are scaled according to the degree of harm

This system rests on the presumption of **the legitimacy of government and the overall sociopolitical order**, and has **human judgment involved at every step**

Areas of Concern — Ethical Frameworks

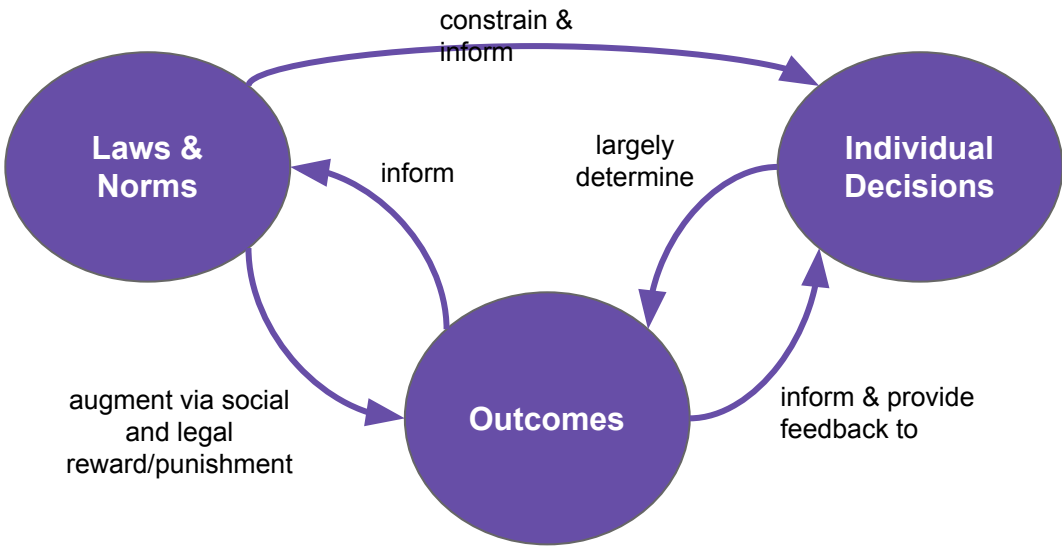
What should we be evaluating to guide the production of ethical technology?

- **Laws & Norms** (at-fault rules)
  - Deontology, Contractualism, Contractarianism
- **Individual decisions** (standards of reasonable care)
  - Virtue ethics, Rule Utilitarianism
- **Outcomes** (minimizing & repairing damages)
  - Consequentialism, Utilitarianism
- **Authority** (who can administer punishment & distribute remuneration)
  - Social Contract Theory

Areas of Concern — In Social Context

How do we understand these loci **within our current sociopolitical context?**

- Who writes the laws, makes impactful decisions, or decides what a “good” outcome is?
- How do ethical norms reflect, and how are they influenced by, power relationships?
- How can being aware of these influences guide our ethics?



Consequentialism

Ethics as an Optimization Problem

*"it is the **greatest happiness** of the **greatest number** that is the measure of right and wrong."*

Jeremy Bentham (1748–1832)

Consequentialism Basics

Consequentialism is the stance that whether an act is ethically right or wrong should be determined by its **consequences.**

Therefore, consequentialism reduces the question of **rightness of an act** to the question of the **goodness of a state of affairs.**

This puts actions on a **continuum of rightness**

- How this translates to moral **obligation** and moral **permission** is an independent question

Formalization

- Let $a \in A$ denote an action and $w \in W$ denote a possible world
  - Worlds w are related by a "betterness" relation $>$
  - World-action pairs (w, a) are related by a "greater rightness" relation $>$.
- Let $T : W \times A \to W$ be the world's transition dynamics.

Consequentialism stipulates that for any world w and actions a_1, a_2,

$$(\mathrm{w}, a_1) > (\mathrm{w}, a_2) \textbf{ if and only if } T(w, a_1) > T(w, a_2).$$

Note: the meaning of this identity **depends on how you formalize *W***, i.e., what gets to count as part of the "outcome"

- If each w includes the history of all actions taken, this could be trivial
- Therefore, consequentialism is best regarded as a **framework to guide one's thinking** rather than a hard constraint on ethical theory

Values

What makes a state of affairs **good**? → That it fulfills desirable **values**.

We can aggregate over individuals:

- **Pleasure vs. pain:** *"the greatest happiness of the greatest number"*
- **Preference satisfaction:** fulfillment of personal interests
- **Quality-Adjusted Life Years:** maximize quality & quantity of life
- **Restitution:** past wrongs to an individual are repaired or compensated for

Or assess global properties:

- **Equality:** *e.g.*, that everyone is afforded equal respect and concern
- **Equity:** that disparities in material condition are minimized

A value V may be formalized as a scalar function on worlds $V : W \to \mathbb{R}$.

Resolving Value Conflicts

- What if values conflict? *e.g.*,
  - Equity ↔ Equality
  - Equality ↔ Pleasure
- Consequentialism requires an **ordering** on $\mathbf{V} = (V_1(w), V_2(w), V_3(w), \ldots)$
- **Pareto efficiency:** no single value V_i can be increased without decreasing another value V_j.
- **Theorem** (Negishi, 1960)*: every **pareto-efficient** allocation of value is the maximum of some **linear combination** of the values

$$\sum_i \alpha_i \cdot V_i(w).$$

  - There's no avoiding the relative weighting of values!

Dealing with Uncertain Outcomes

Moral responsibility is for the outcome, but it can't always be foreseen

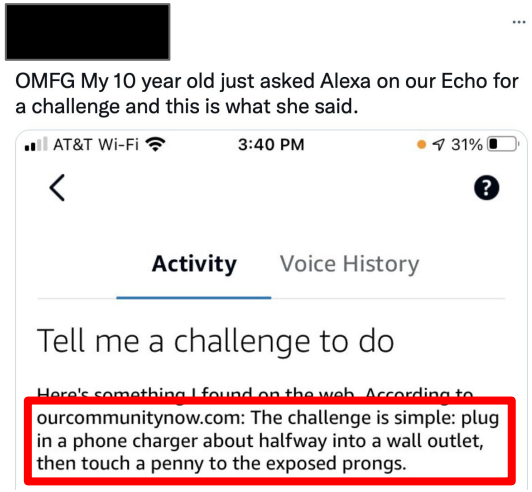
- **Information gathering** about **foreseeable outcomes** is ethically important
  - William K. Clifford, 1877: *The Ethics of Belief*

von Neumann–Morgenstern (VNM) Utility Theorem:

Any agent obeying certain axioms of rationality must be behaving in a way that* ***maximizes the expected utility of some utility function****.*

$$\text{action}(w) = \text{argmax}_a \sum_{w'} p_T(w' \mid a, w) \cdot \sum_i \alpha_i \cdot V_i(w')$$

- This requires estimating uncertainties correctly!



Microsoft Tay (2016–2016)

https://PollEv.com/multiple_choice_polls/q1Vu2qlafxqzjcUMQBFww/respond

You receive a **$1 billion** grant for the development of **large language models**.

- You think these models have huge potential to revolutionize NLP.
- You also worry deeply about bias and misuse of such models.
- You believe a competing lab is working on LLMs, but you have an advantage.

What do you do?

a) 📖 Train your model and open-source it for the research community to scrutinize freely

b) 🔒 Train the model but keep it behind an API where you limit its users and can carefully observe and control how it is used

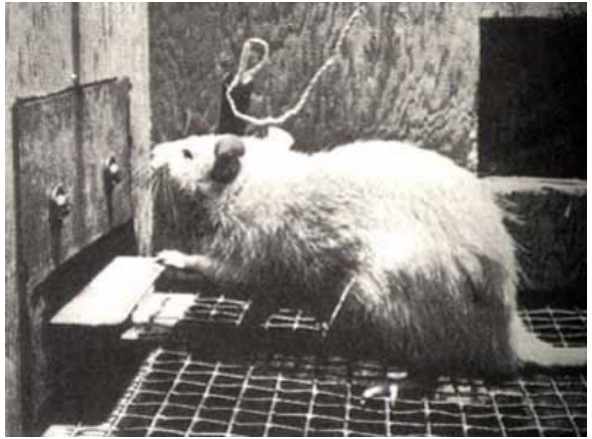
c) 🕵️ Instead of training models, fund secret operations to sabotage competitors and delay the development of LLMs as much as possible

d) ❓ Other.

How do you measure utility?

Revealed Preferences

"Wireheading"



MONEY

How much money would someone have to pay you to quit Facebook? Study suggests it's $1,000

Published Mon, Jan 7 2019•9:35 AM EST • Updated Mon, Jan 7 2019•10:24 AM EST

Yoni Blumberg
@YONIBLUM

SHARE

A facebook display during the China International Import Expo (CIIE), at the National Exhibition and Convention Center in Shanghai, China, November 5, 2018. Aly Song | Reuters

What should our goals be?

- Deciding what's good for people seems bad.
- Egging them on to keep doing what they're doing (just more of it) also seems bad.

How do we **help people live their best lives, whatever that means for them?**

- Can your technology do this? How?

Autonomy & Authority

Moral Plurality and Social Contracts

The good people of your nation are suffering.

- ***Infant mortality*** *rate is high.*
- ***Women are bound*** *to their homes.*
- *Men are subjected to* ***manual labor****.*
- ***Medical care*** *is poor.*
- *People* ***aren't free*** *to pursue their passions.*

Let us send you our mechanized agricultural equipment and free the hands of your people to higher-skilled work, building out your infrastructure and economy, so you may compete on the world stage and raise your standard of living.

No, thank you.

- *Our society has lasted* ***thousands of years****.*
- *We are* ***well established*** *in our ways of life.*
- *Relieving our men of hard labor would* ***rob them of their mettle****.*
- *Relieving our mothers of their children would* ***rob them of their purpose****.*
- *Globalizing trade would* ***rob us of our unique culture, dilute our language and food, and threaten the shared beliefs and values*** *that keep our society together and make it what it is.*

The good people of your language community are suffering.

- ***Few internet pages and services** are available in your language.*
- *Your people are not competitive on the **international labor market.***
- *Your people are missing out on the benefits of **advancements in language technology**.*

We can build you machine translation systems. We have enough data to get started, but we need native speakers to help construct parallel corpora.

No, thank you.

- *Our language has been around **thousands of years**, and suits our needs just fine.*
- *Developing these resources would make our society **more legible to surveillance**.*
- *It will **expose our workers** to labor competition from non-natives.*
- *And it will **dilute our language and culture** as future generations will lack an economic incentive to preserve it.*

Navigating Moral Plurality

- There may be no rational, objective way to determine the **"good"**
- Engelhardt: the only "secularly" ethical society is one which allows **moral strangers to coexist**
- This demands two levels of morality
  - A **secular morality** for moral strangers
  - A **community morality** for moral friends
- Therefore, secular ethics must be based on the **principle of permission**
  - Authority and punishment must derive from this

Moral Friends

Moral Strangers

H. Tristram Engelhardt Jr., *Foundations of Bioethics*, 2nd Ed., 1996.

Deontology

- **Deontology:** ethics where the criteria for right and wrong are not in the consequences of an action, but the **rules that it follows**
- Common themes of deontological systems:
  - Requiring consent / preserving autonomy
  - Making and adhering to contracts
  - Applying universal principles to all individuals

The Plight of the Libertarian Fish Farmers

- 100 independent fish farms operate on the same lake, each producing $100/day of fish under good conditions
- Fish farming pollutes the lake: one farm's pollution reduces each farm's yield by $1/day
- But a farm can stop polluting if it runs a $30/day filtration system

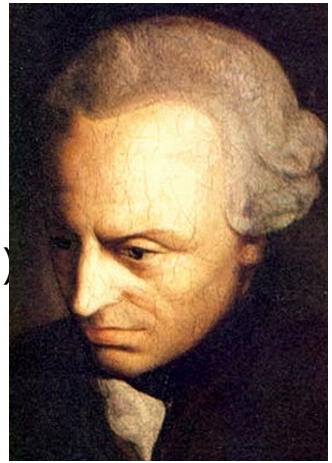
Do the farmers have a moral obligation to filter their pollution?

https://PollEv.com/multiple_choice_polls/1jjRIJkXt0Wvujjh65cvb/respond

Example adapted from Scott Alexander, 2014.

Kantian Fish Farming

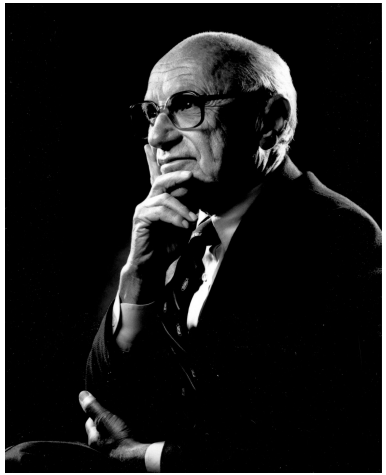
- The **categorical imperative:** *"Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."*
- The categorical imperative allows for moral reasoning via **reductio ad absurdum:**
  - *Should I run my filter?*
    *→ No, **since it reduces my profit*** (maximize)
    *→ if **everyone** did the same, all incomes would be $0* (universalize)
    *→ which **defeats** my original reasoning* (contradict)
    → ⊥
- Kant believed a single set of **universal moral laws** could be derived on such reasoning alone



Immanuel Kant
(1724–1804)

Contractual Fish Farming

- **Contractualism** (T.M. Scanlon, 1998): moral actions are those which obey rules that “nobody can reasonably reject”
- **Contractarianism:** Moral rightness/wrongness must be grounded in **mutual agreement** between persons
  - Some view this as a practical way to maximize utility without stipulating good/bad for other people
  - But contracts differ in theory & practice:
    - Who reads Terms of Service?
    - What about monopolies?
  - Still doesn’t solve the fish farmers’ problem!



Milton Friedman
(1912–2006)

Social–Contractual Fish Farming

Social Contract Theory: the authority of a state derives from an implicit **contract** with the people:

- The state is granted **authority** to exert coercive power (punish, collect taxes, etc.)
- …as long as it fulfills its end of the bargain, to:
  - Establish security (Thomas Hobbes),
  - Protect individual rights (John Locke), or
  - Enact the will of the people (Jean-Jacques Rousseau)
- **You do not "opt out"** of the social contract — only rebel, vote for changes, etc.
  - Fish-farming filter inspections required by law

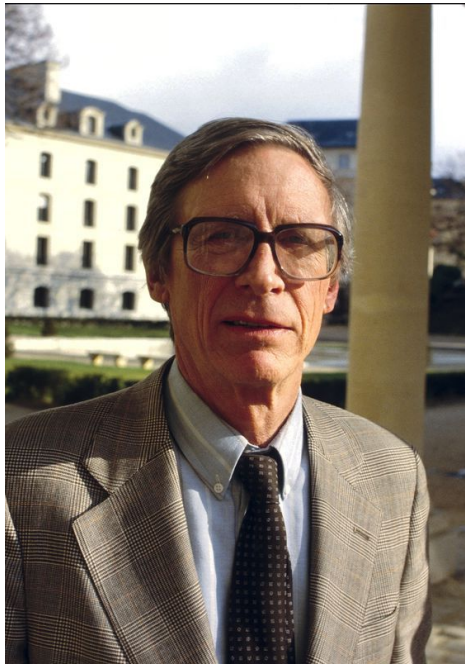
Cover of *Leviathan*, by Thomas Hobbes (1588–1679)

Liberal Fish Farming

Political Liberalism: social contract theory based on rights, welfare, and justice

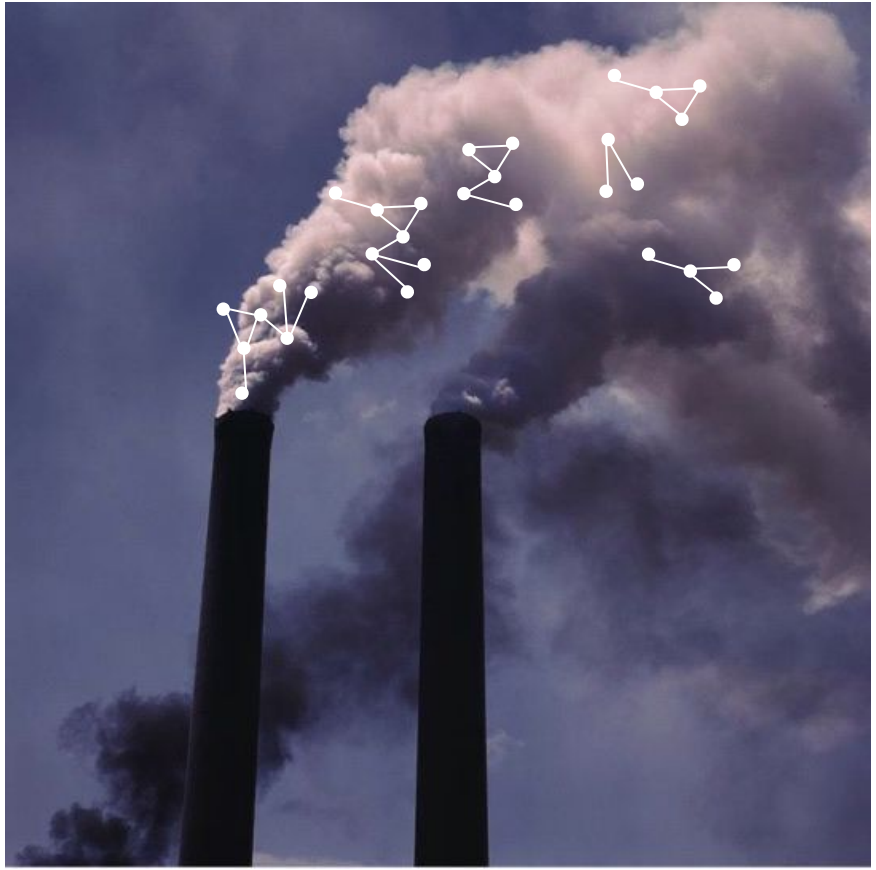
Rawls's **Veil of Ignorance:** the rules of the social contract should be determined based on what anyone would accept *without knowledge* of their place in society

- Prioritizes the needs of the **worst-off**
- Allows for inequality **in cases where the worst-off would accept it**



John Rawls (1921–2002)

Can a river consent? Does it have rights? Interests?



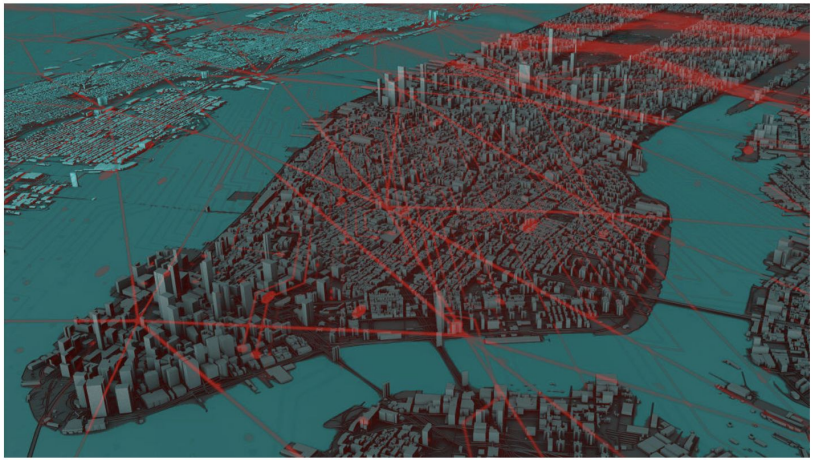
What obligations do we owe the public square of data?

Can AI be Regulated?

11-26-19 | POV

The first effort to regulate AI was a spectacular failure

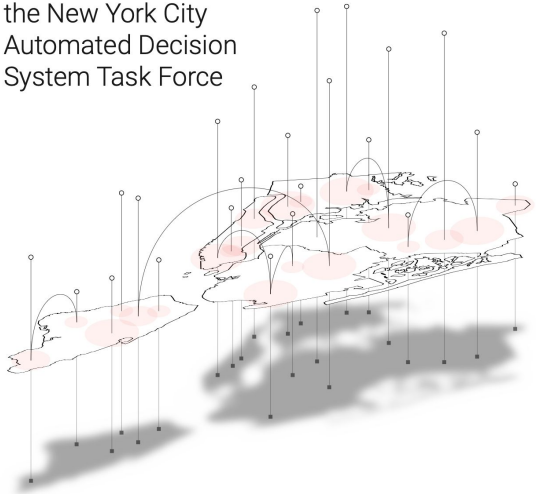
Albert Fox Cahn participated in New York City's process to understand how automated decision systems are impacting its citizens. But, he writes, the historic, well-intentioned initiative went "horribly wrong."



[Source illustrations: yucelyilmaz/iStock; Naeblys/iStock]

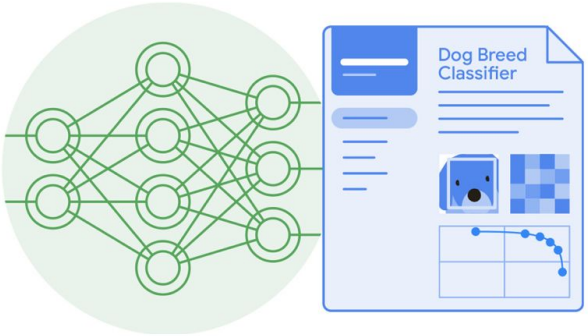
CONFRONTING BLACK BOXES

A Shadow Report of the New York City Automated Decision System Task Force

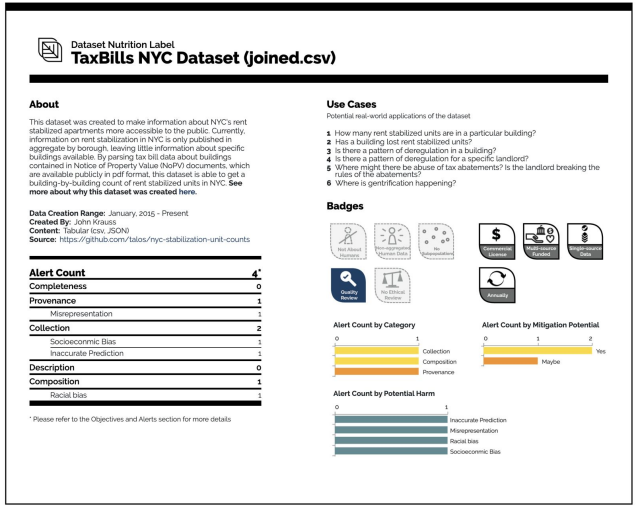


DECEMBER 2019

Transparency & Understanding



Model Cards
(Mitchell *et al.*, 2018)



Dataset Nutrition Label
TaxBills NYC Dataset (joined.csv)

About

This dataset was created to make information about NYC's rent stabilized apartments more accessible to the public. Currently, information on rent stabilization in NYC is only published in aggregate by borough, leaving little information about specific buildings available. By parsing tax bill data about buildings contained in Notice of Property Value (NoPV) documents, which are available publicly in pdf format, this dataset is able to get a building-by-building count of rent stabilized units in NYC. **See more about why this dataset was created here.**

Data Creation Range: January, 2015 - Present
Created By: John Krauss
Content: Tabular (csv, JSON)
Source: https://github.com/talos/nyc-stabilization-unit-counts

Alert Count	4*
Completeness	0
Provenance	1
Misrepresentation	1
Collection	2
Socioeconmic Bias	1
Inaccurate Prediction	1
Description	0
Composition	1
Racial bias	1

* Please refer to the Objectives and Alerts section for more details

Use Cases

Potential real-world applications of the dataset

1. How many rent stabilized units are in a particular building?
2. Has a building lost rent stabilized units?
3. Is there a pattern of deregulation in a building?
4. Is there a pattern of deregulation for a specific landlord?
5. Where might there be abuse of tax abatements? Is the landlord breaking the rules of the abatements?
6. Where is gentrification happening?

Badges

Not About Humans | Non-aggregate Human Data | No Preprocessing | Commercial License | Independently Applied | Single-source data | Quality Review | No Ethical Review | Annually

Alert Count by Category

0 — 1
Collection
Composition
Provenance

Alert Count by Mitigation Potential

0 — 1 — 2
Yes
Maybe

Alert Count by Potential Harm

0 — 1
Inaccurate Prediction
Misrepresentation
Racial bias
Socioeconmic Bias

Dataset Nutrition Labels
(Holland *et al.*, 2018)

Critical Theory

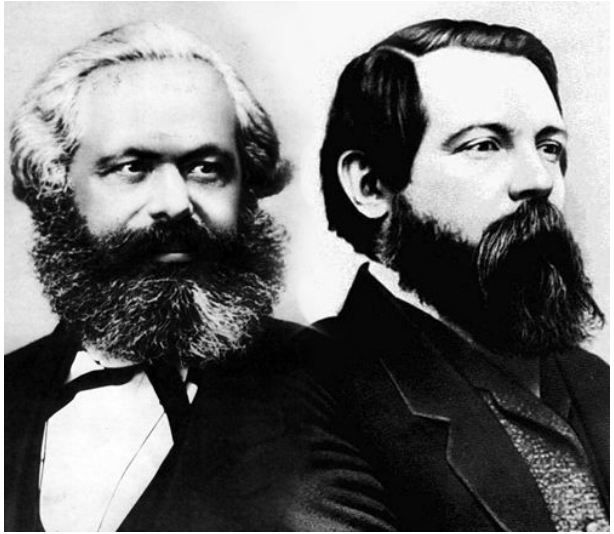
Class, Identity, and Power

Historical Materialism

"The history of all hitherto existing society is the history of class struggles."

— Karl Marx, *The Communist Manifesto*, 1848

- Marx and Engels argued that the development of society could be explained by
  - the **organization of the methods of economic production**, which resulted in
  - **material conditions** experienced by groups,
  - which could be identified with the interests of their **socioeconomic class.**
- We can analyze society in a way that treats classes as **agents acting in their interests**.



Karl Marx (1818–1883) and Friedrich Engels (1820–1895)

Axes of Power & Domination

Broadly focused on **identifying and challenging hierarchies of power**

- Marxism — economic class
- Feminism — gender & sex
- Critical race theory — race
- Postcolonialism — colonialism & imperialism

Intersectionality: understanding the discrimination in the intersections between power structures

Frantz Fanon
(1925–1961)

Kimberlé Crenshaw

Constructed Categories & Identities

"One is not born, but rather becomes, woman."

— Simone de Beavoir, *The Second Sex*

"A man who has a language consequently possesses the world expressed and implied by that language."

— Frantz Fanon, *Black Skin, White Masks*

- Social and linguistic categories are constructed by those in power, and **often serve the interests** of those in power.

Simone de Beauvoir
(1908–1986)

Relation to Ethical Theory

Consequentialist elements:

- ***Who*** *is benefiting?* ***Who*** *is being harmed?*
- *Will there be* ***disparate impact****?*

Non-consequentialist elements:

- *Do I have the authority to* ***categorize people****? How will the categories be* ***used****?*
- *How does automating this function* ***concentrate power****?*

These two aspects may be in tension:

- *How can I quantify & prevent bias* ***without imposing categories on people****?*
- *How can I prevent disparate outcomes without* ***unduly concentrating power to myself****?*

Ethical Tensions

- **Bias & Discrimination:**
  - Understanding disparate impact requires categorizing people; who decides the categories?
  - Mitigating bias requires resolving value conflicts, but who decides the weights?
- **Hate Speech & Toxicity:**
  - Does censoring online hate speech centralize power in the media platform?
  - Hate speech classifiers are biased when trained without extra context — but is incorporating that extra context a form of surveillance?
- **Misinformation**
  - We want to prevent its spread, but who has the authority to arbitrate truth?
- **Privacy**
  - Preserving privacy requires smoothing over individuals — will preserving people's rights produce a worse result?

End