# FSDL 2022

# Ethics

Charles Frye

October 3, 2022

# Agenda

00

## Context

What do we talk about when we talk about ethics?

01

## Tech Ethics

What has ethics meant in tech so far?

02

## ML Ethics

What does ethics mean in ML right now?

03

## AI Ethics

What might ethics mean in an AI future?

00

# Context

# All ethics lectures are wrong, but some are useful.

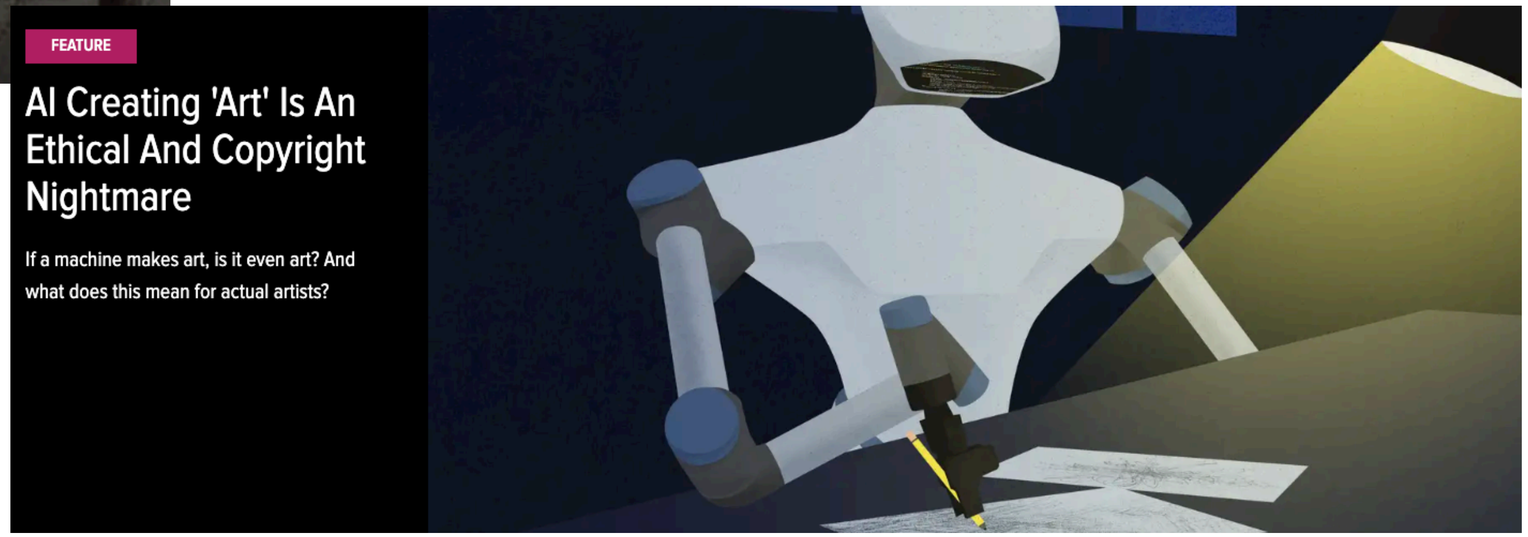
- The approach of this lecture
- Three themes:
  - Alignment
  - Trade-Offs
  - Humility

We'll approach ethics on the basis of concrete cases.

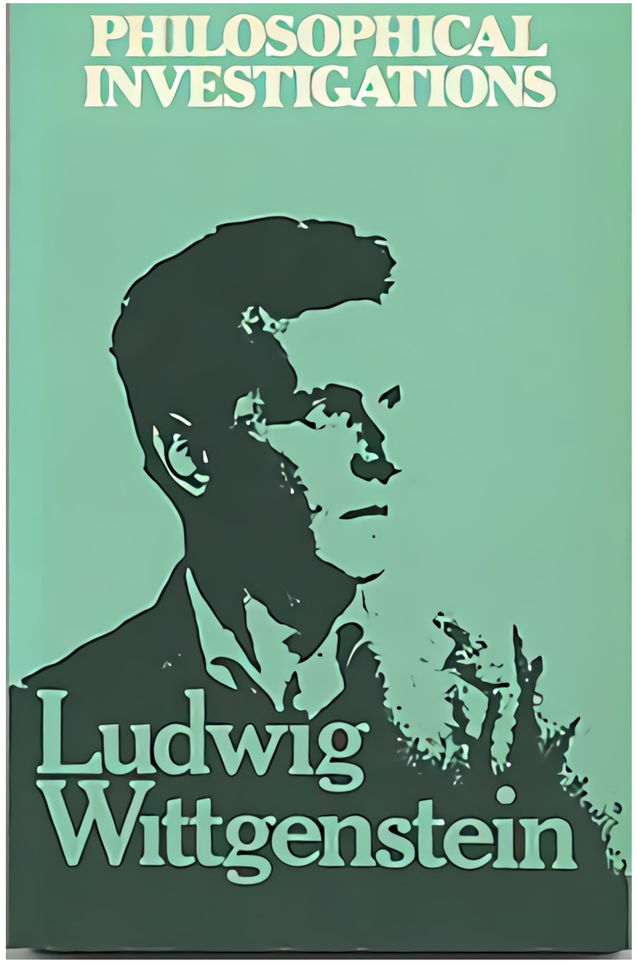
# We'll talk about specific cases where actions have lead to claims and counter-claims of "ethical" or "unethical" behavior.

**A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.**

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel
October 17, 2016 at 5:00 a.m. EDT



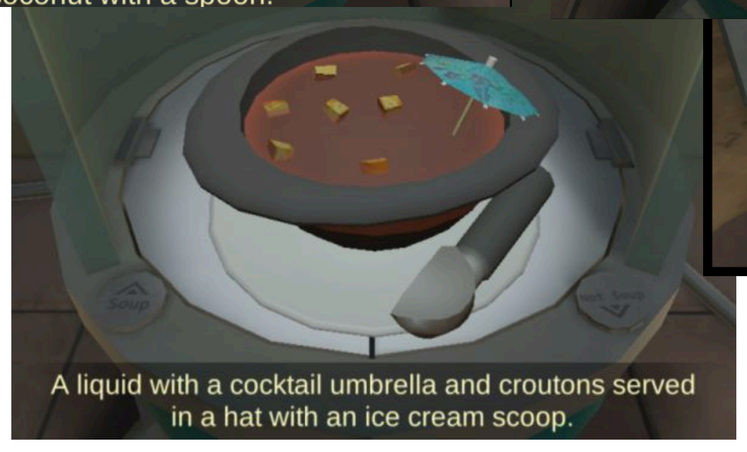
# What is ethics?

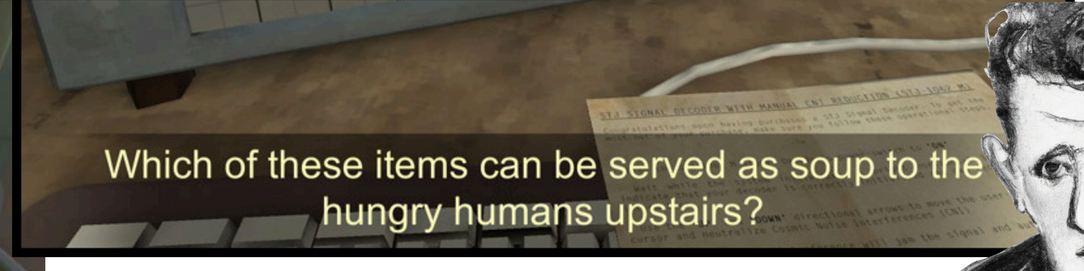


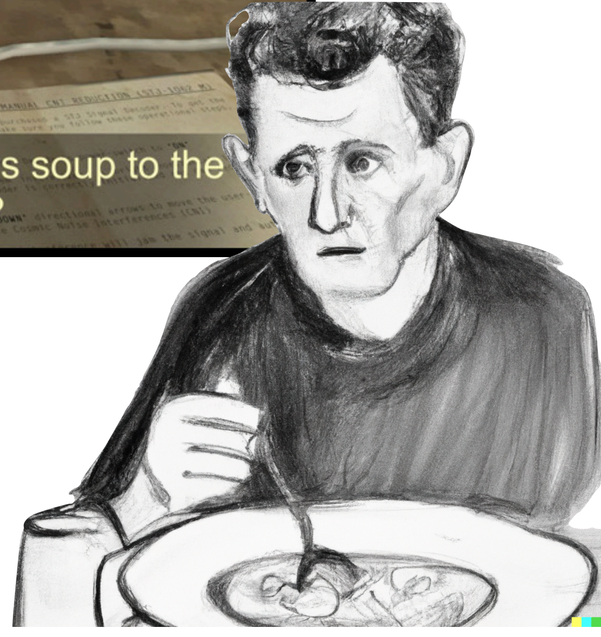

"The meaning of a word is its use in the language."

Image upsampled by the "AI Image Enlarger" at https://bigjpg.com/

# Something Something Soup Something



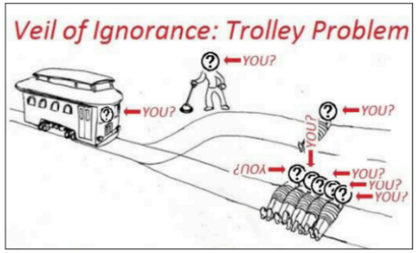
https://soup.gua-le-ni.com/

# We won't talk about ethical schools or "trolley" problems.

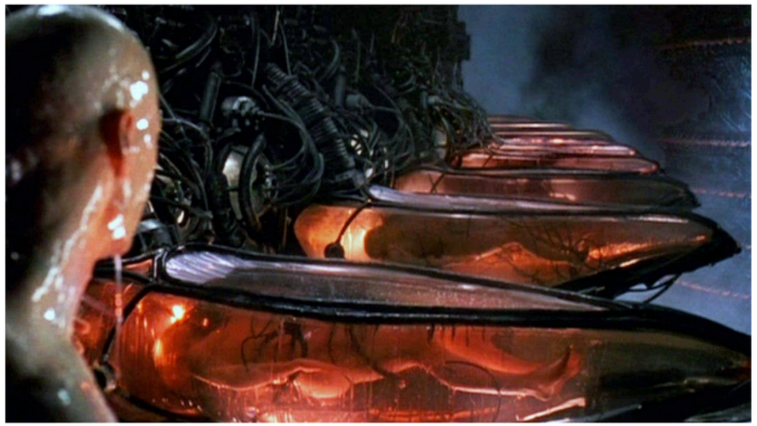
YOU ARE ON AN ASTEROID CAREENING through the cosmos. Aboard the asteroid with you are nine hundred highly-skilled physicians, who have been working on developing a revolutionary medication that will cure every disease in the known universe. The asteroid's current trajectory is taking it straight toward the Planet of Orphans, where all intergalactic civilizations have dumped their unwanted offspring, of which there are now 100 trillion, all living, breathing, and mewling. If you detonate the asteroid, all of the doctors will die, along with the hope for curing every disease in the universe. If you do not detonate the asteroid, the doctors will have time to develop the cure and send it hurtling toward the Healing Planet before you crash into and destroy the Planet of Orphans. Thus you face the crucial question: *how useful is this hypothetical for illuminating moral truths?*

https://www.currentaffairs.org/2017/11/the-trolley-problem-will-tell-you-nothing-useful-about-morality

# For another point of view, see Sergey's 2021 lecture.

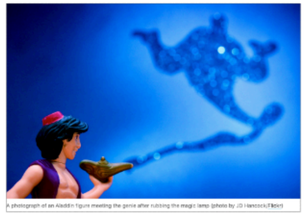


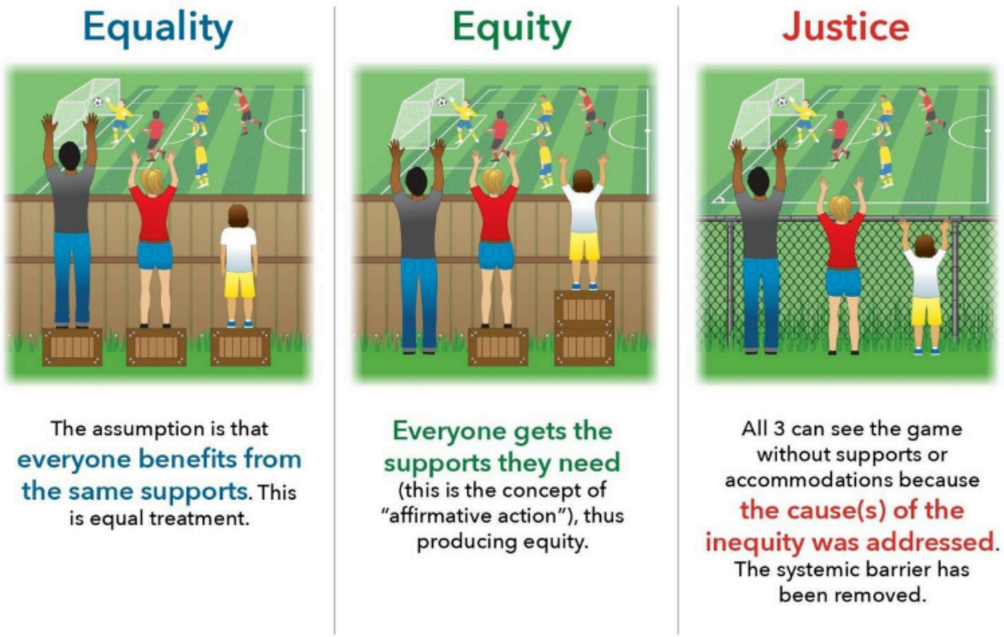

## Replacing humans entirely



## Alignment

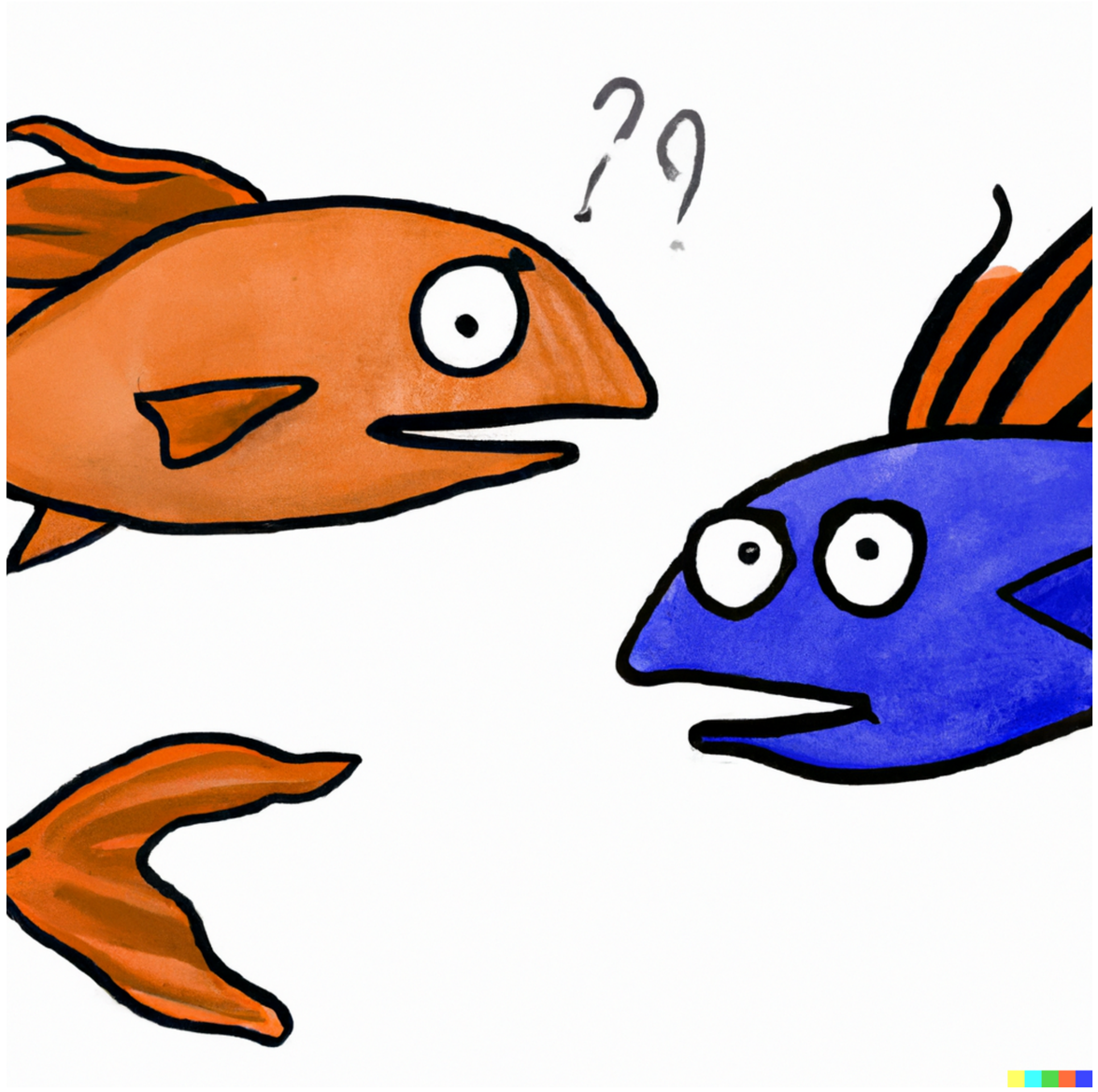
- "Paperclip maximizer": AGI given the goal of producing paperclips eventually turns every atom in space into paperclips
- Old lesson: establishing and communicating our goals and values is hard, and technology amplifies the difficulty






https://fullstackdeeplearning.com/spring2021/lecture-9/

# A useful theme from 2021: "what is water?"



# That is: "what is vanishing into the background?"

- Relying on prominent cases risks replicating social biases.
  Whose ethical claims are amplified? Which stories travel?

- Case-based reasoning can hide the principles that are actually in operation, making them disappear like water.

**No approach is perfect, and ethics is deeply personal.**

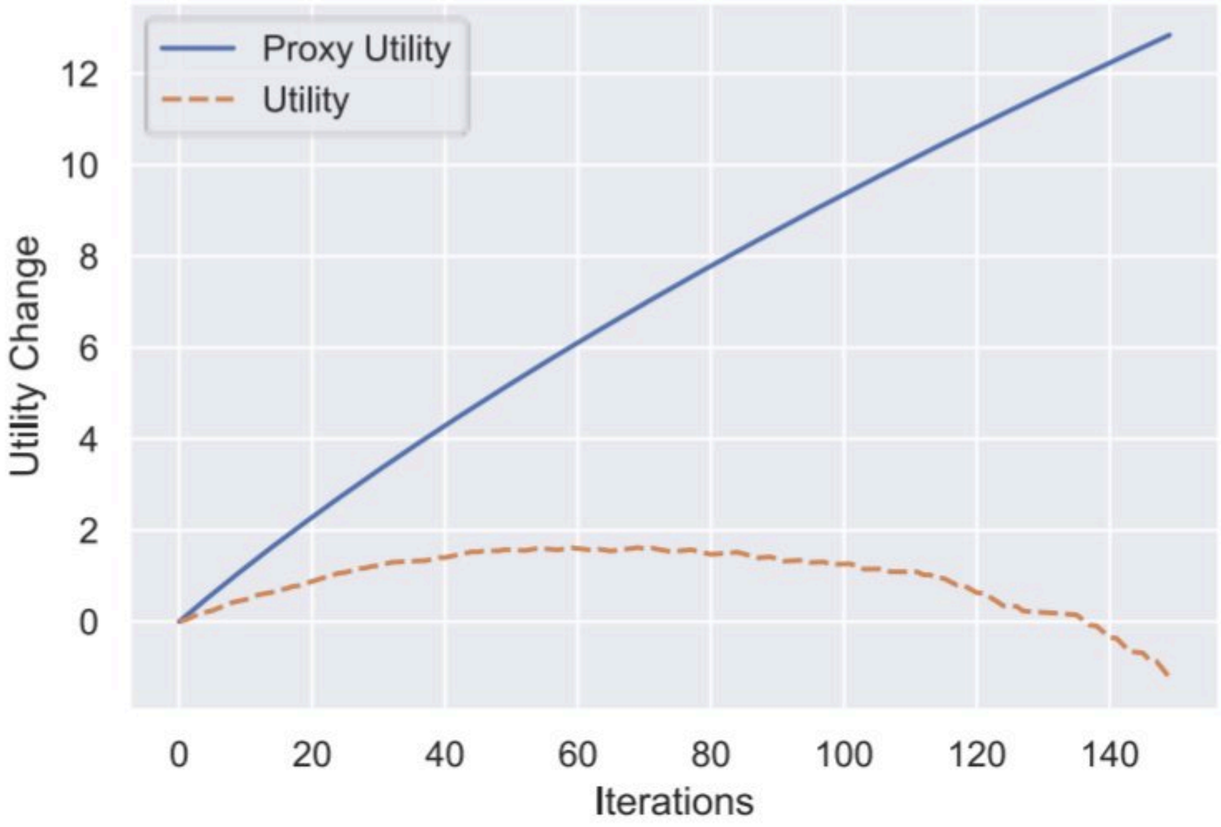
We'll see three themes repeatedly.

- 👯 **Alignment** 👯
  we know what we want, but we don't get it.

- ⚖️ **Trade-Offs** ⚖️
  different entities want different things.

- 🤦 **Humility** 🤦
  we don't know what we want or how to get it.

# Alignment and the proxy problem.




## Loosely:

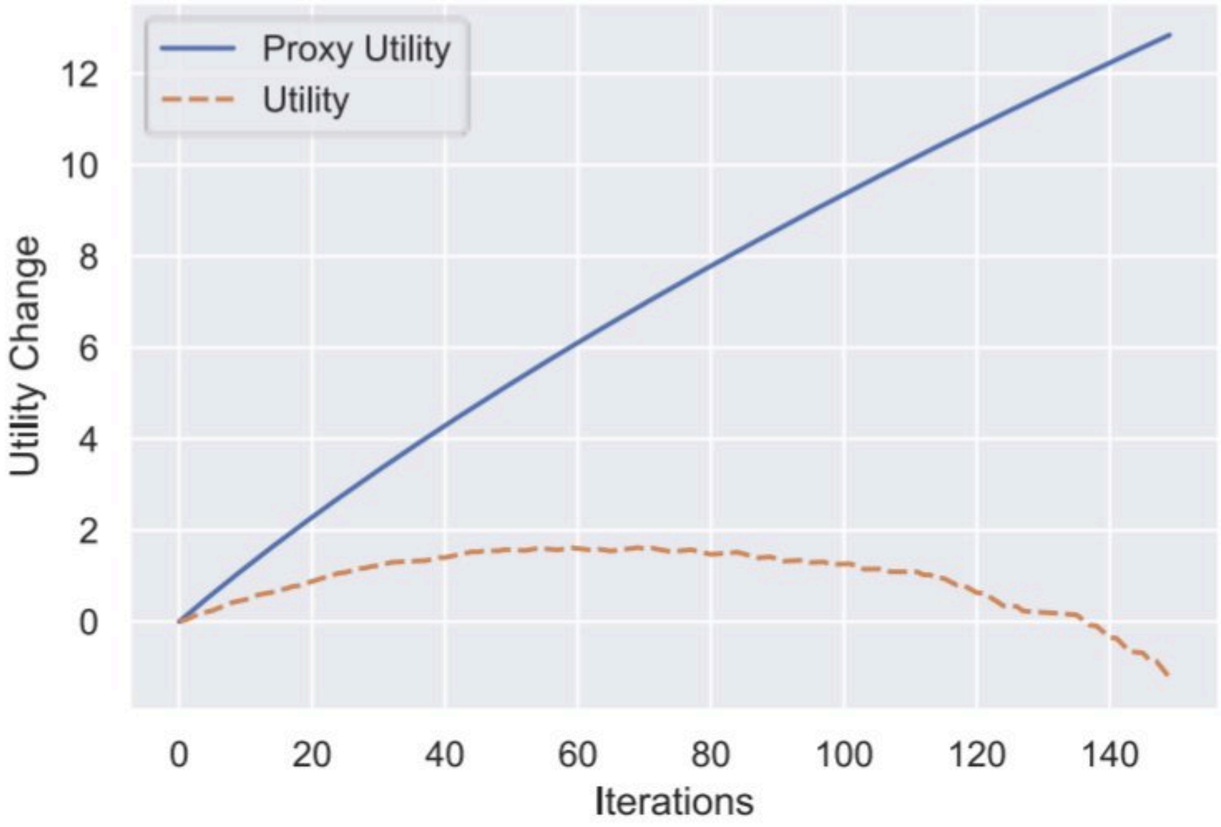
If a proxy can be improved arbitrarily by disregarding true utility,

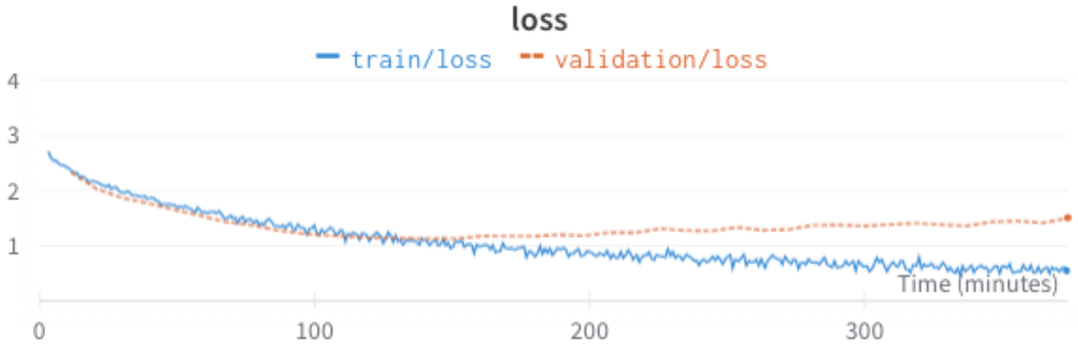
and if utility can get arbitrarily low,

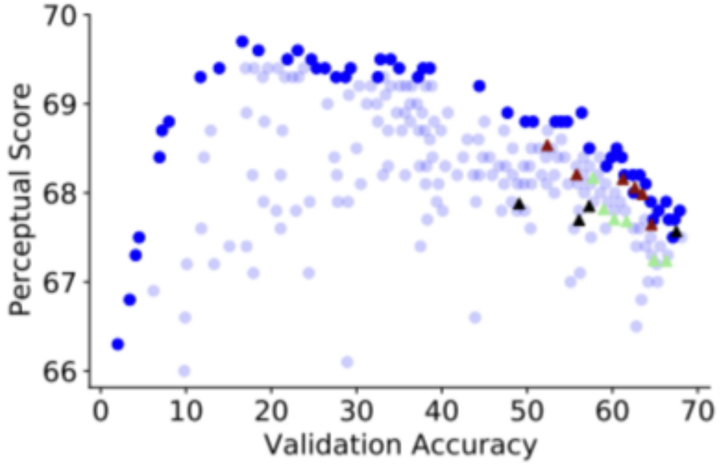
any agent that maximizes the proxy will inevitably decrease utility.

https://arxiv.org/abs/2102.03896

# Proxy problems abound.



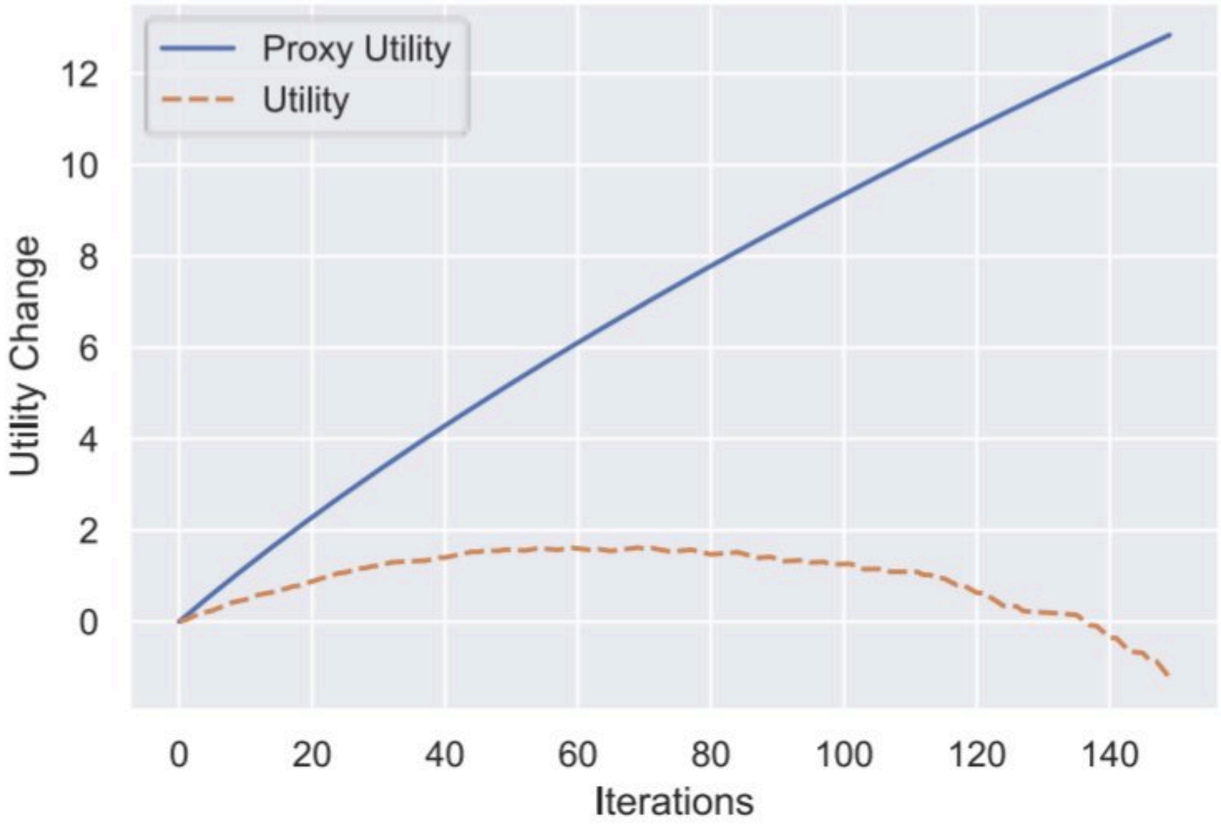




https://openreview.net/forum?id=qrGKGZZvH0

# 👯 Alignment

# Proxy problems abound.




A graphic example of the difficulties which this creates for the economy is provided by an enterprise in Tambov which was making obsolete, though adequate, machines for vulcanising tyres. This gave it a 1968 sales plan of 8,600,000 roubles. It was proposed that in 1969 it should switch over to a new automatic line which vulcanises at much greater speed and with considerable economy of labour. Prices, however, are such that sales in 1969 would be only 5,000,000 roubles if the new machine were produced.

'Isn't national economic efficiency taken into account in determining the prices of new products?'

'The price of new chemical equipment depends primarily on its weight.' (Belyaev [1968].)

# 👯 Alignment

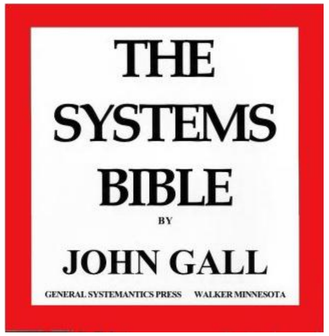
https://skeptics.stackexchange.com/questions/22375/did-a-soviet-nail-factory-produce-useless-nails-to-improve-metrics

# Proxy problems abound.

**Systemantics Quotes**
@SysQuotes

THE INFORMATION YOU HAVE IS NOT THAT INFORMATION YOU WANT.

51 8:57 AM - Feb 20, 2022




**Systemantics Quotes**
@SysQuotes

THE INFORMATION YOU WANT IS NOT THE INFORMATION YOU NEED.

80 8:57 AM - Feb 21, 2022

## On the folly of rewarding A, while hoping for B

Steven Kerr

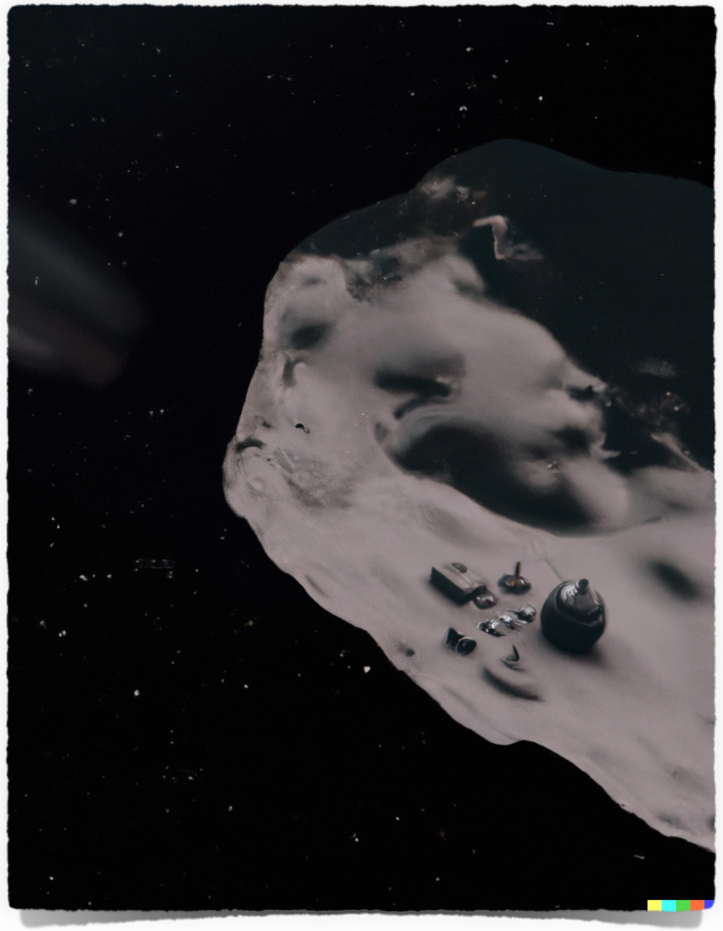
Whether dealing with monkeys, rats, or human beings, it is hardly controversial to state that most organisms seek information concerning what activities are rewarded, and then seek to do (or at least pretend to do) those things, often to the virtual exclusion of activities not rewarded. The extent to which this occurs of course will depend on the perceived attractiveness of the rewards offered, but neither operant nor expectancy theorists would quarrel with the essence of this notion.

Nevertheless, numerous examples exist of reward systems that are fouled up in that the types of behavior rewarded are those which the rewarder is trying to discourage, while the behavior desired is not being rewarded at all.

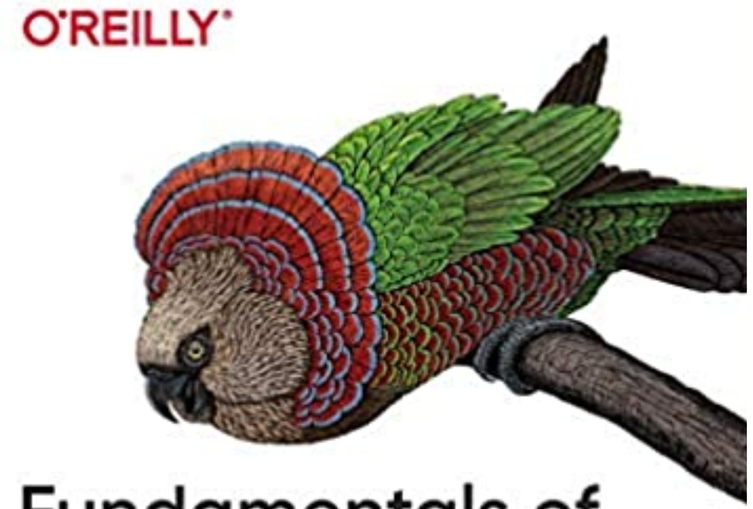
https://www.ou.edu/russell/UGcomp/Kerr.pdf

# Alignment

# ⚖️ Trade-offs arise between stakeholders.

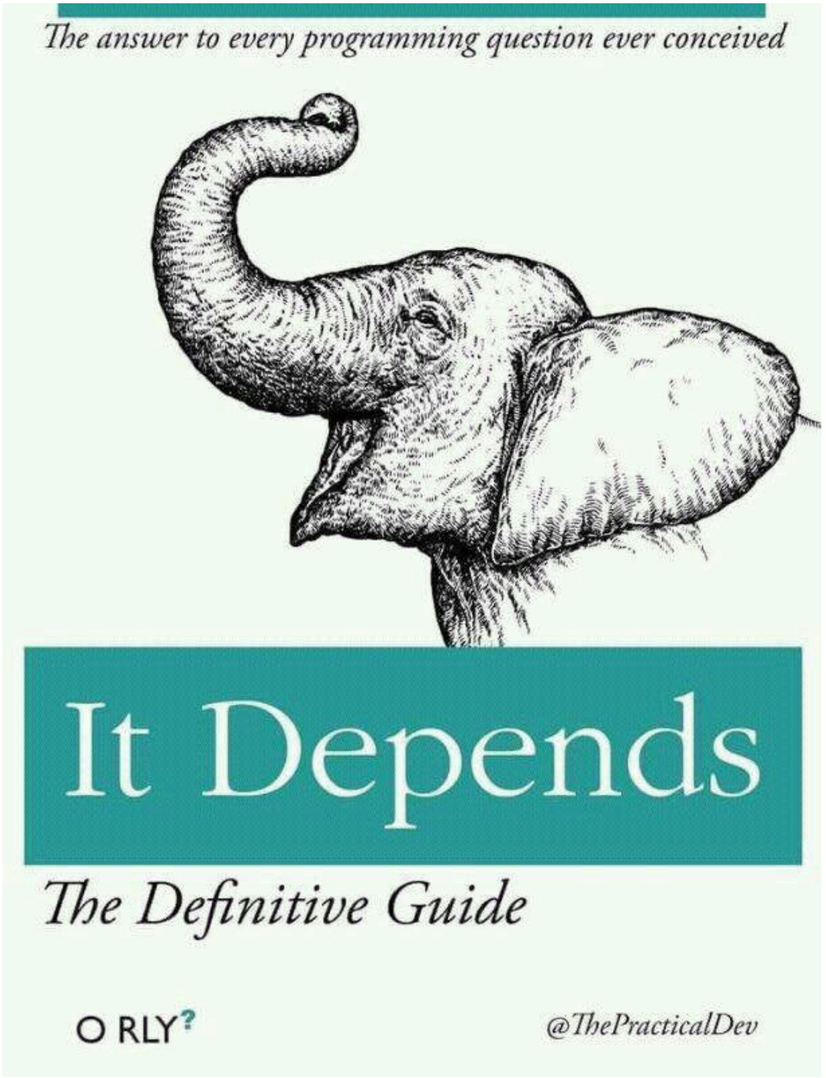


# We accept trade-offs in engineering.



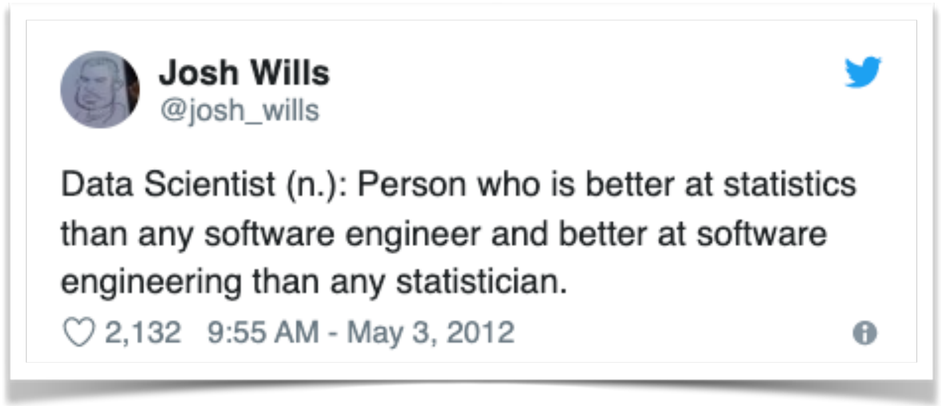

While the scope of software architecture is almost impossibly broad, unifying elements do exist. The authors have first and foremost learned the *First Law of Software Architecture* by constantly stumbling across it:

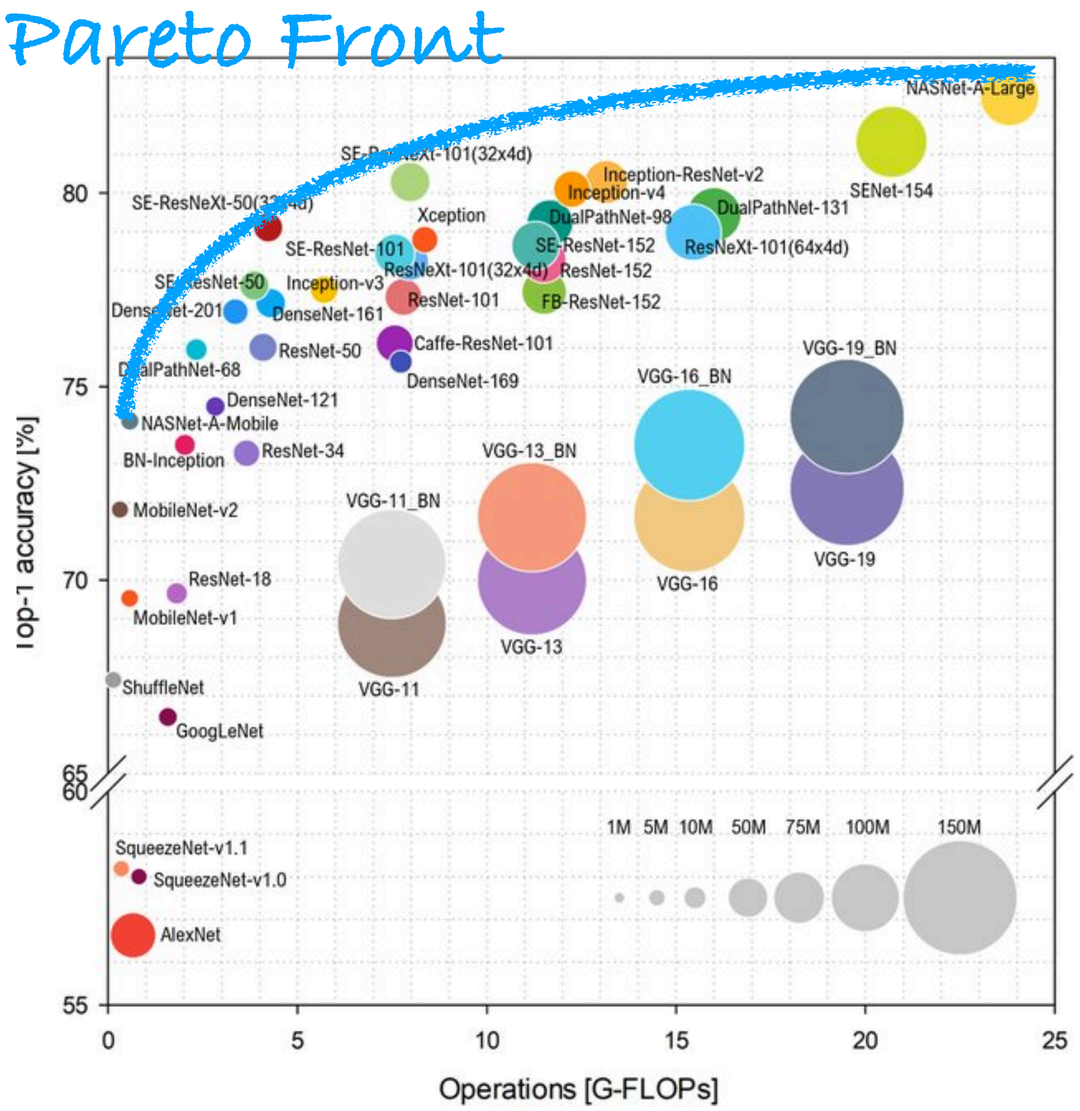
> Everything in software architecture is a trade-off.
>
> —First Law of Software Architecture




⚖️ Trade-Offs

# We accept trade-offs in engineering.



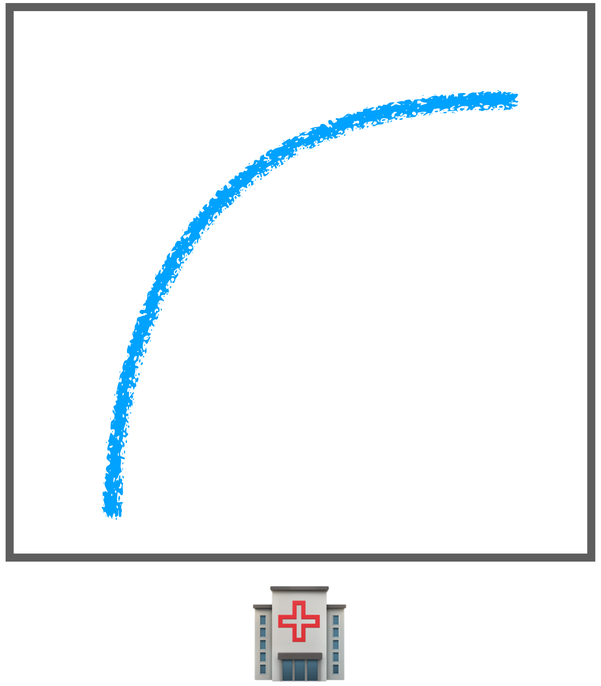




## ⚖️ Trade-Offs

# Ethical trade-offs are hardmode.

- Hard to identify, hard to quantify (👯 proxy problem)

- Once measured, where on the Pareto front do we fall?
  - How many current orphans 👼 do we trade for future health 🏥?




- Many questions outside of our expertise.

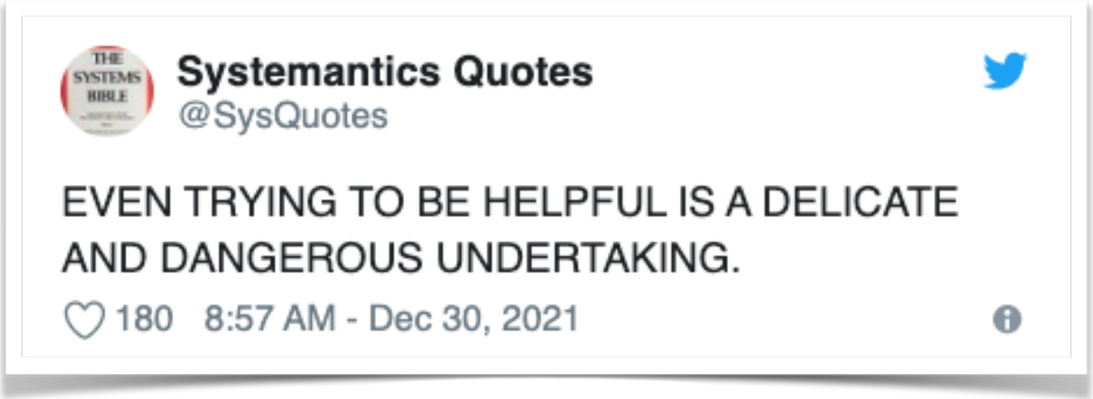
## ⚖️ Trade-Offs

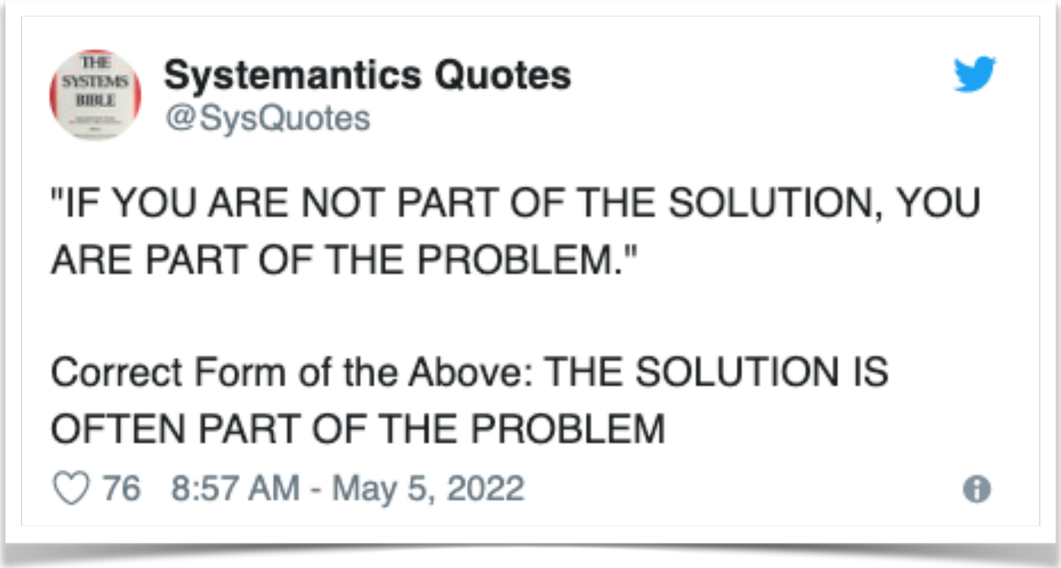
# Humility is the appropriate response.

- Most engineers do not explicitly train these skills.

- Many engineers and managers in tech in fact constitutionally prefer optimizing single metrics ("number go up", no ⚖️) that are not proxies (no 👯).

- Don't forget humility when intervening!



# User orientation undergirds each theme.

- Feedback maintains 👯 alignment between our system and users.
- ⚖️ Trade-offs are resolved in consultation with users.
- 🤷 Humility means we actually listen to users.

01

# Tech Ethics

# The tech industry cannot afford to ignore ethics.

- Tech's "public approval ratings" are in decline
- We can learn from nearby industries
- Some contemporary topics

# The past decade of tech has been plagued by scandal.

**Ben Collins**
@oneunderscore__

Quick thread:

I want you all to meet Vladimir Bondarenko.

He's a blogger from Kiev who really hates the Ukrainian government.

He also doesn't exist, according to Facebook.

He's an invention of a Russian troll farm targeting Ukraine. His face was made by AI.

65.7K 10:29 AM - Feb 28, 2022

The New York Times

TECHNOLOGY

## India permanently bans TikTok and 58 other Chinese apps

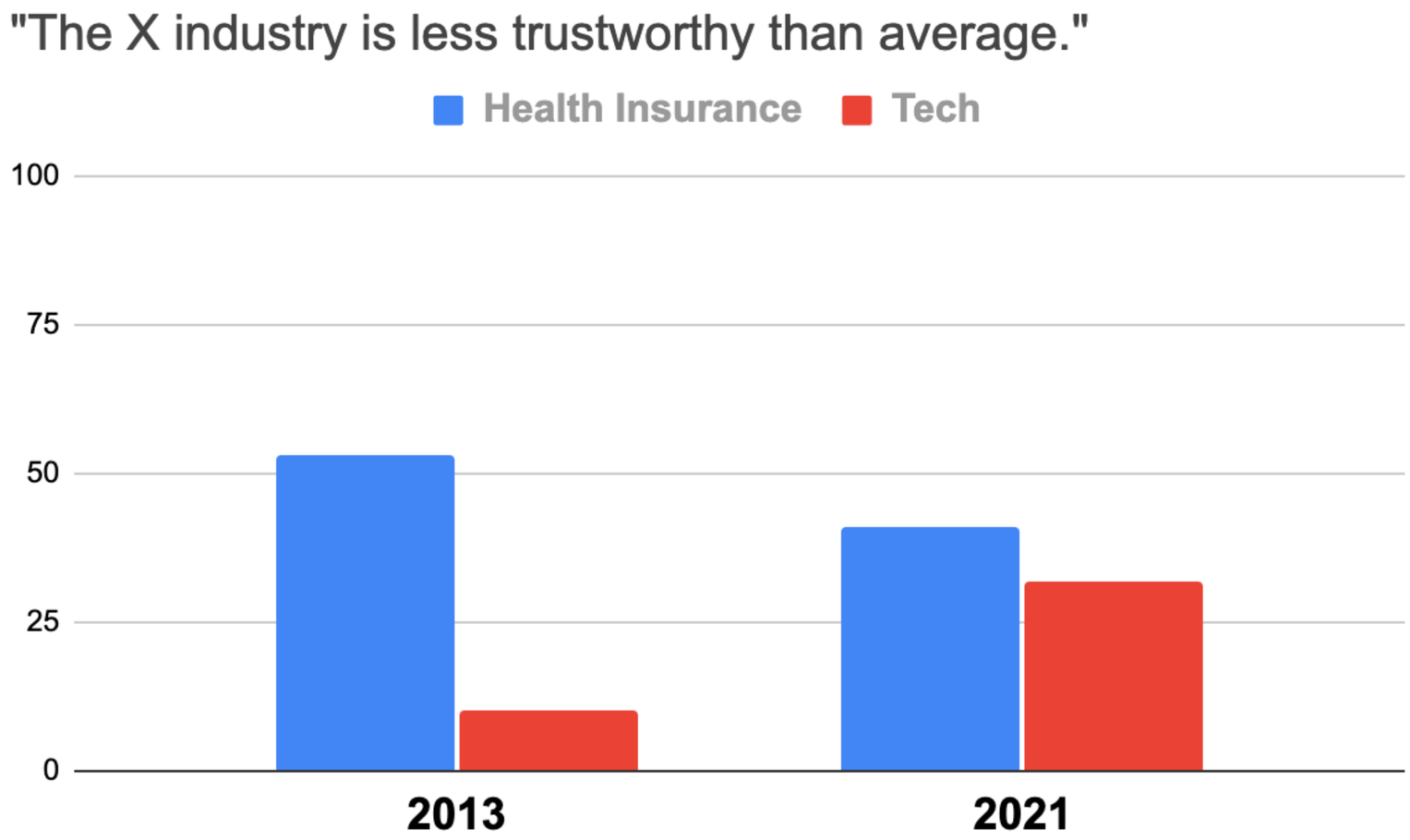
Government skeptical of developers' privacy and security assurances

*Google Will Not Renew Pentagon Contract That Upset Employees*

the facebook files

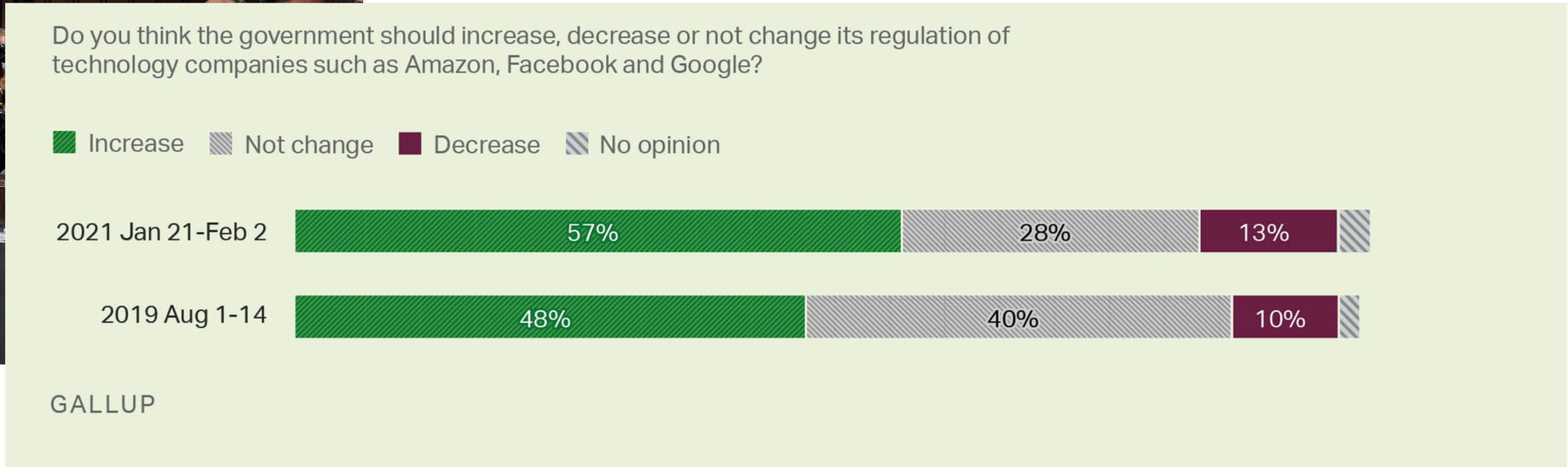
## Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show

# Distrust of tech companies rose sharply in the last decade.




https://pac.org/public-affairs-pulse-survey-2021

# Distrust leads to regulatory intervention.




In 2011, the fraction wanting more regulation of tech was more like 10-20%.

## 👯 Alignment

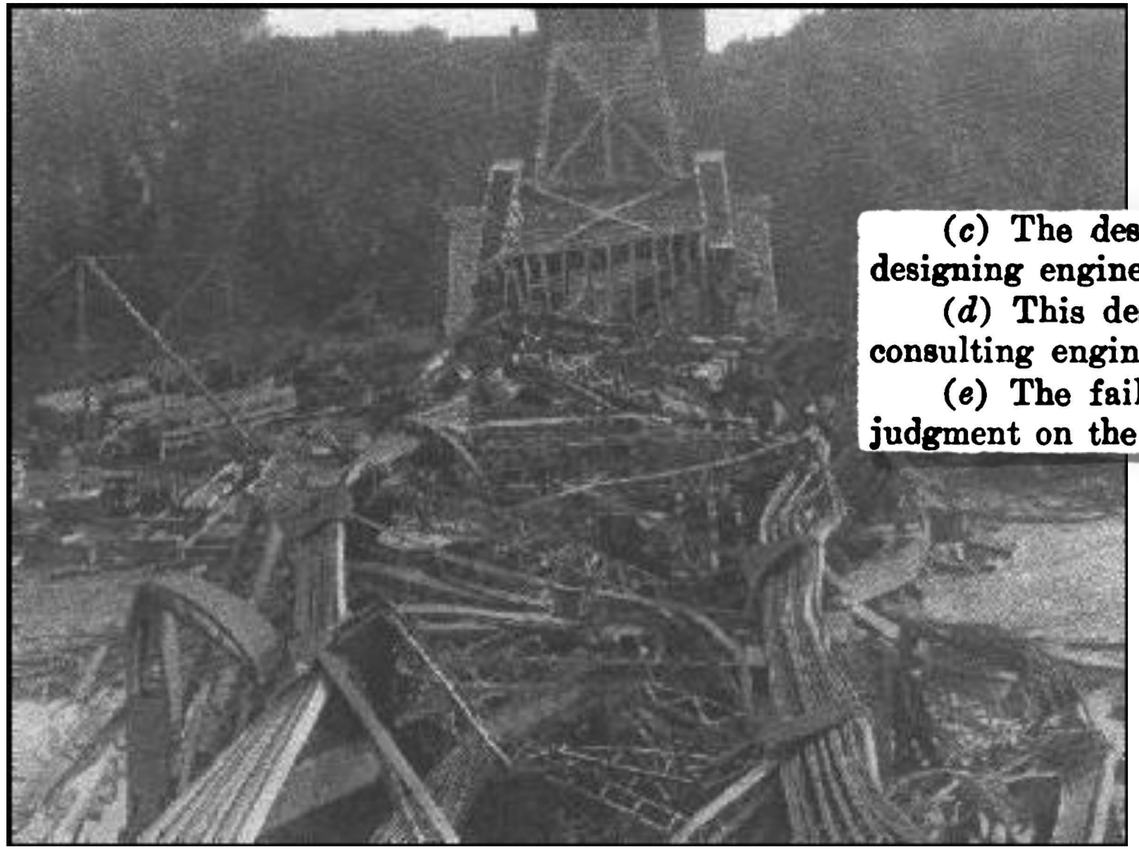
https://news.gallup.com/poll/329666/views-big-tech-worsen-public-wants-regulation.aspx

# We can learn from nearby fields.

- Canadian engineers:
  the bridge, the ring,
  and the calling.

- Human subjects research:
  from Nuremberg
  to Menlo Park.

# In 1907, the Quebec Bridge collapsed, killing 75.



(*c*) The design of the chords that failed was made by Mr. P. L. Szlapka, the designing engineer of the Phœnix Bridge Company.

(*d*) This design was examined and officially approved by Mr. Theodore Cooper, consulting engineer of the Quebec Bridge and Railway Company.

(*e*) The failure cannot be attributed directly to any cause other than errors in judgment on the part of these two engineers.

Parliamentary inquiry placed blame on the engineers.

Iron, that, of the best of my knowledge and power, I will not henceforward suffer or pass, or be privy to the passing of, Bad Workmanship or Faulty Material in aught that concerns my works before mankind as an Engineer, or in my dealings with my own Soul before my Maker.

FOR MY ASSURED FAILURES and derelictions, I ask pardon beforehand of my betters and my equals in my Calling here assembled; praying that in the hour of my temptations,

To this day, many Canadian engineers participate in rituals that impress upon them the weight of their responsibility.

# Humility

# Research is also subject to regulation.

- WWII-era medical experiments on unwilling human subjects by totalitarian regimes lead to Nuremberg Code (1947).
  - Through Helsinki Declaration (1964), this is still the cornerstone for laws on human subjects research worldwide.
- US: 1973 National Research Act requires informed consent.
  - Infliction of hepatitis on mentally disabled children in New York
  - Non-treatment of syphilis in Black men at Tuskegee



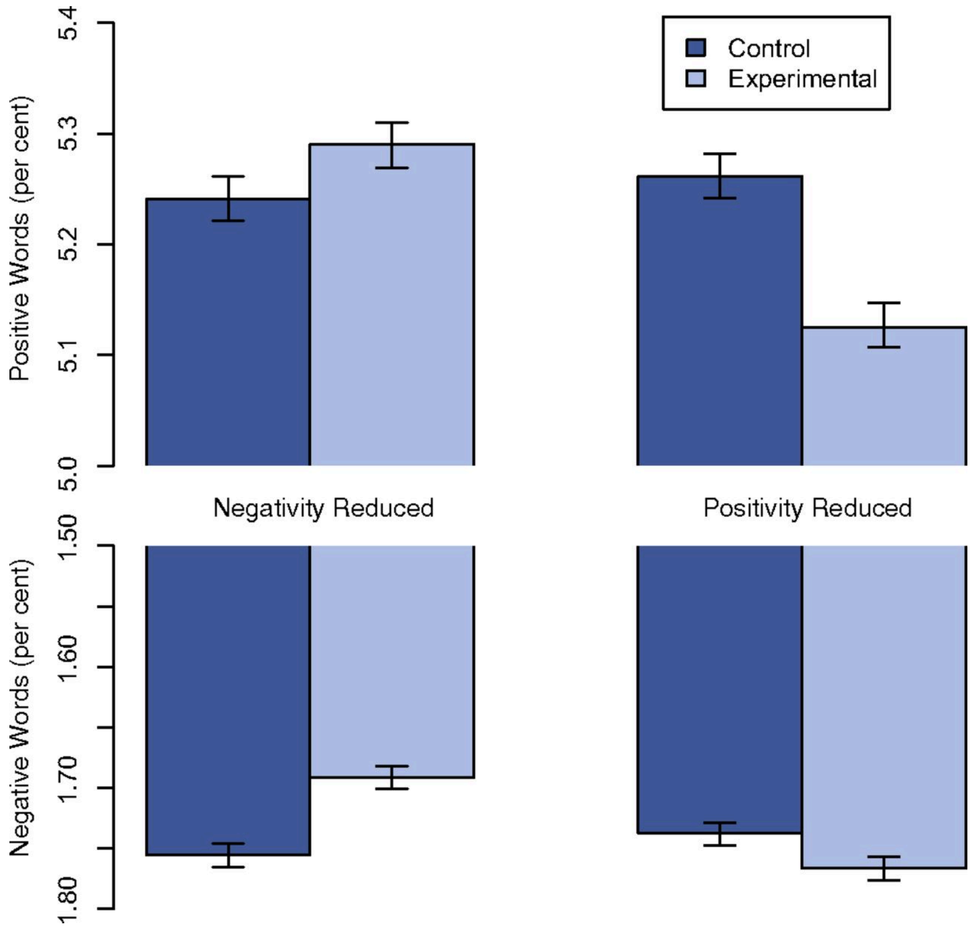
https://www.niehs.nih.gov/research/resources/bioethics/timeline/index.cfm

# We are also researchers and are expected to adhere to the same principles.

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |

## Experimental evidence of massive-scale emotional contagion through social networks

Questions have been raised about the principles of informed consent and opportunity to opt out in connection with the research in this paper. The authors noted in their paper, "[The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors




https://www.pnas.org/doi/full/10.1073/pnas.1320040111

We need a professional code of ethics for software.

# Some concrete ethical dilemmas

- Dark patterns and user-hostile design
- Carbon emissions from energy-intensive work

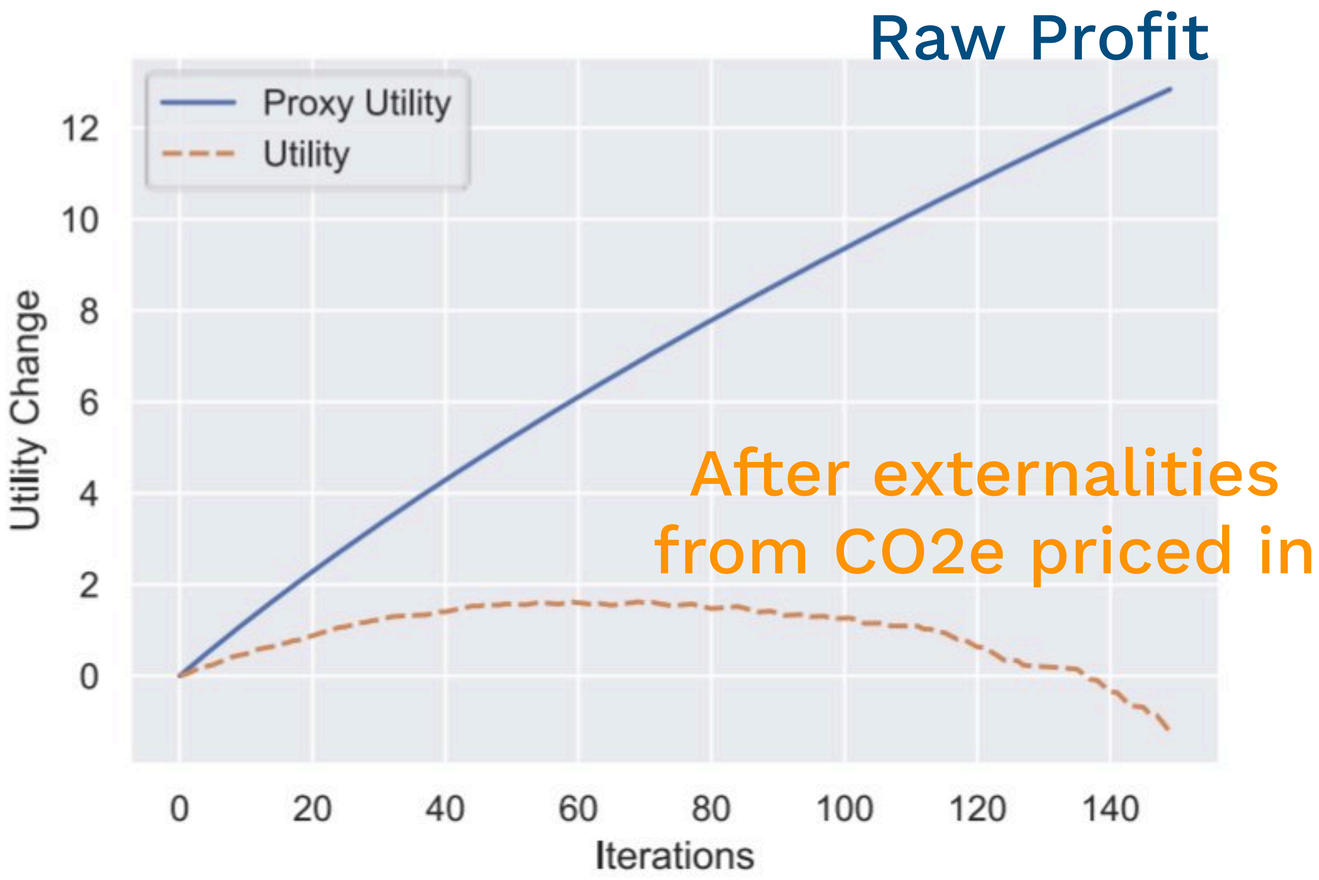
# Carbon emissions scale with cost.

# Anthropogenic climate change from CO2 emissions raises ethical concerns.

George Jetson sues the present for melting the ice caps.
*Harvey Birdman, Attorney at Law*

## ⚖️ Trade-Offs

# One view: CO2e is a classic alignment problem.




## Alignment

# Compute jobs require power, which correlates with CO2e.

| Model | Hardware | Power (W) | Hours | kWh·PUE | $CO_2e$ | Cloud compute cost |
|---|---|---|---|---|---|---|
| $T2T_{base}$ | P100x8 | 1415.78 | 12 | 27 | 26 | \$41–\$140 |
| $T2T_{big}$ | P100x8 | 1515.43 | 84 | 201 | 192 | \$289–\$981 |
| ELMo | P100x3 | 517.66 | 336 | 275 | 262 | \$433–\$1472 |
| $BERT_{base}$ | V100x64 | 12,041.51 | 79 | 1507 | 1438 | \$3751–\$12,571 |
| $BERT_{base}$ | TPUv2x16 | — | 96 | — | — | \$2074–\$6912 |
| NAS | P100x8 | 1515.43 | 274,120 | 656,347 | 626,155 | \$942,973–\$3,201,722 |
| NAS | TPUv2x1 | — | 32,623 | — | — | \$44,055–\$146,848 |
| GPT-2 | TPUv3x32 | — | 168 | — | — | \$12,902–\$43,008 |

Table 3: Estimated cost of training a model in terms of $CO_2$ emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

| **Consumption** | **$CO_2e$ (lbs)** |
|---|---|
| Air travel, 1 person, NY↔SF | 1984 |
| Human life, avg, 1 year | 11,023 |
| American life, avg, 1 year | 36,156 |
| Car, avg incl. fuel, 1 lifetime | 126,000 |

| **Training one model (GPU)** | |
|---|---|
| NLP pipeline (parsing, SRL) | 39 |
| w/ tuning & experiments | 78,468 |
| Transformer (big) | 192 |
| w/ neural arch. search | 626,155 |

## But power isn't free.

On US-based cloud infra,
$10 roughly equal to $1 of air travel costs.

The U.S. Environmental Protection Agency (EPA) provides average $CO_2$ produced (in pounds per kilowatt-hour) for power consumed in the U.S. (EPA, 2018), which we use to convert power to estimated $CO_2$ emissions:

$$CO_2e = 0.954 p_t \quad (2)$$

https://aclanthology.org/P19-1355/

# Simply changing cloud regions can reduce emissions.

| region | country | city | impact |
|---|---|---|---|
| ap-south-1 | India | Mumbai | 920.0 |
| ap-southeast-2 | Australia | Sydney | 802.0 |
| ap-east-1 | China | Hong Kong | 702.0 |
| cn-north-1 | China | Beijing | 680.0 |
| eu-west-2 | United Kingdom | London | 623.0 |
| eu-west-1 | Ireland | Dublin | 617.0 |
| eu-central-1 | Germany | Frankfurt am Main | 615.0 |
| us-east-2 | USA | Columbus | 568.2 |
| ap-northeast-2 | South Korea | Seoul | 517.0 |
| ap-northeast-3 | Japan | Osaka | 516.0 |
| ap-southeast-1 | Singapore | Singapore | 419.0 |
| us-east-1 | USA | Ashburn | 367.8 |
| us-west-2 | USA | Portland | 297.6 |
| us-west-1 | USA | San Francisco | 240.6 |
| sa-east-1 | Brazil | Sao Paulo | 205.0 |
| eu-west-3 | France | Paris | 105.0 |
| eu-north-1 | Sweden | Stockholm | 47.0 |
| ca-central-1 | Canada | Montreal | 20.0 |

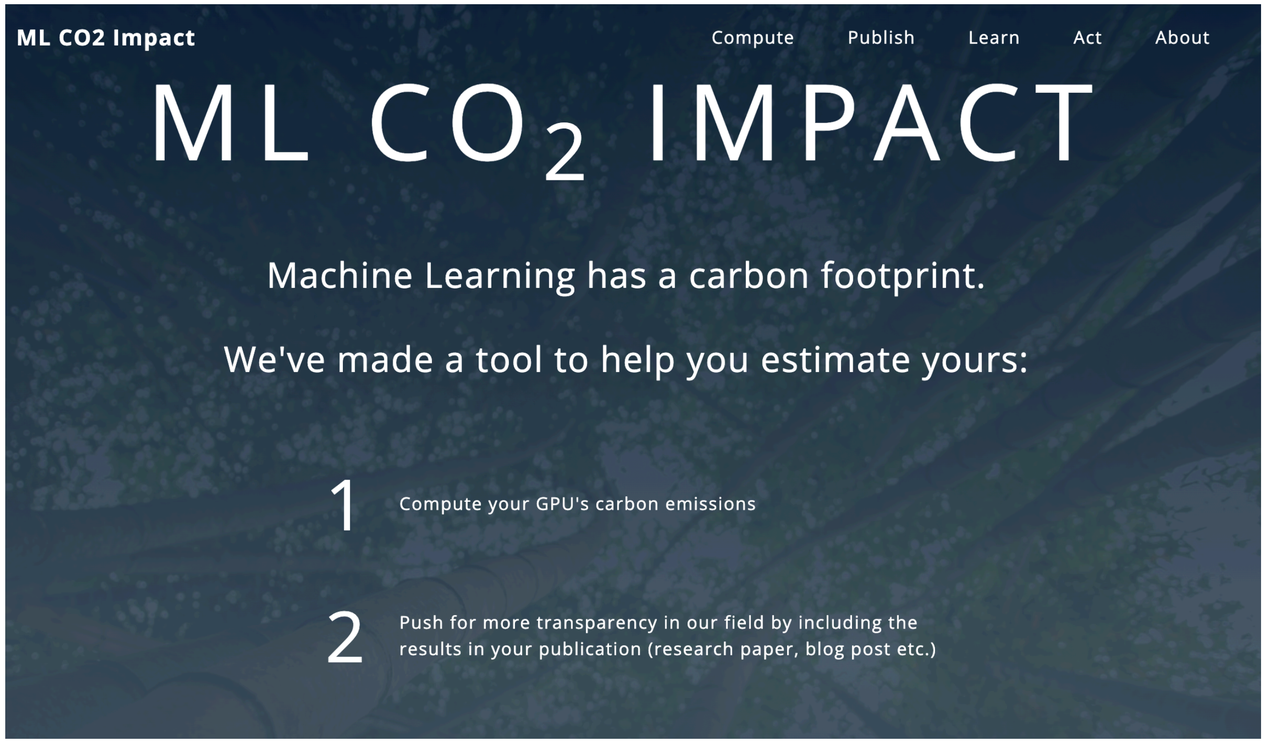
| **Consumption** | **$CO_2e$ (lbs)** |
|---|---|
| Air travel, 1 person, NY↔SF | 1984 |
| Human life, avg, 1 year | 11,023 |
| American life, avg, 1 year | 36,156 |
| Car, avg incl. fuel, 1 lifetime | 126,000 |
| **Training one model (GPU)** | |
| NLP pipeline (parsing, SRL) | 39 |
| w/ tuning & experiments | 78,468 |
| Transformer (big) | 192 |
| w/ neural arch. search | 626,155 |

(US-based cloud infra)

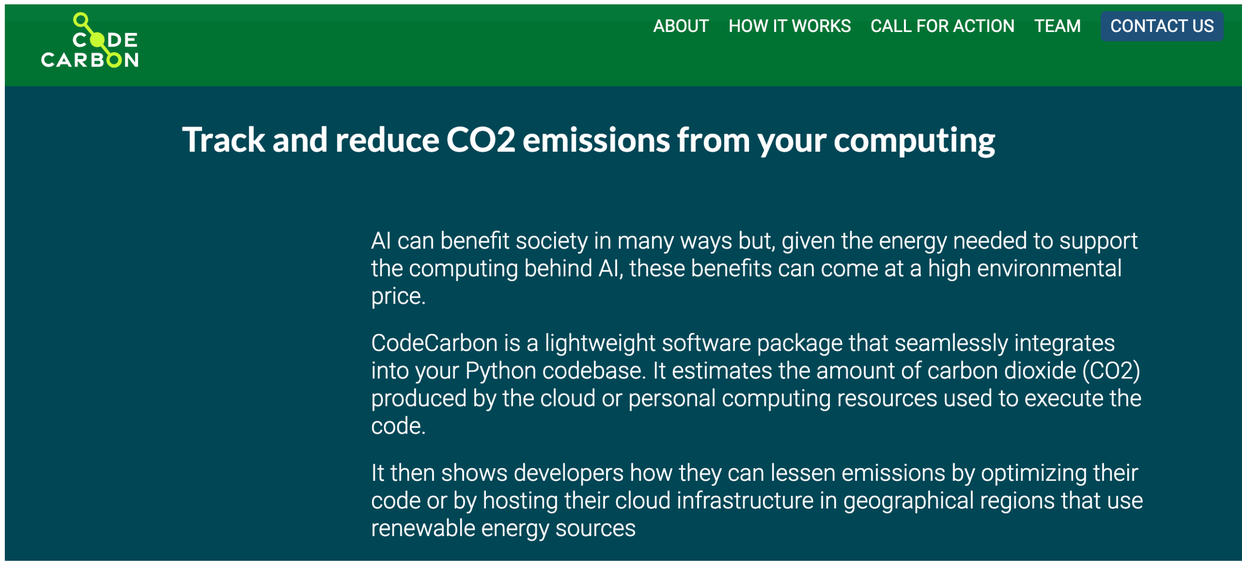
A factor of nearly 50x from most to least CO2e.

From Hugging Face: https://www.youtube.com/watch?v=ftWlj4FBHTg

# New tooling helps!




https://mlco2.github.io/impact/

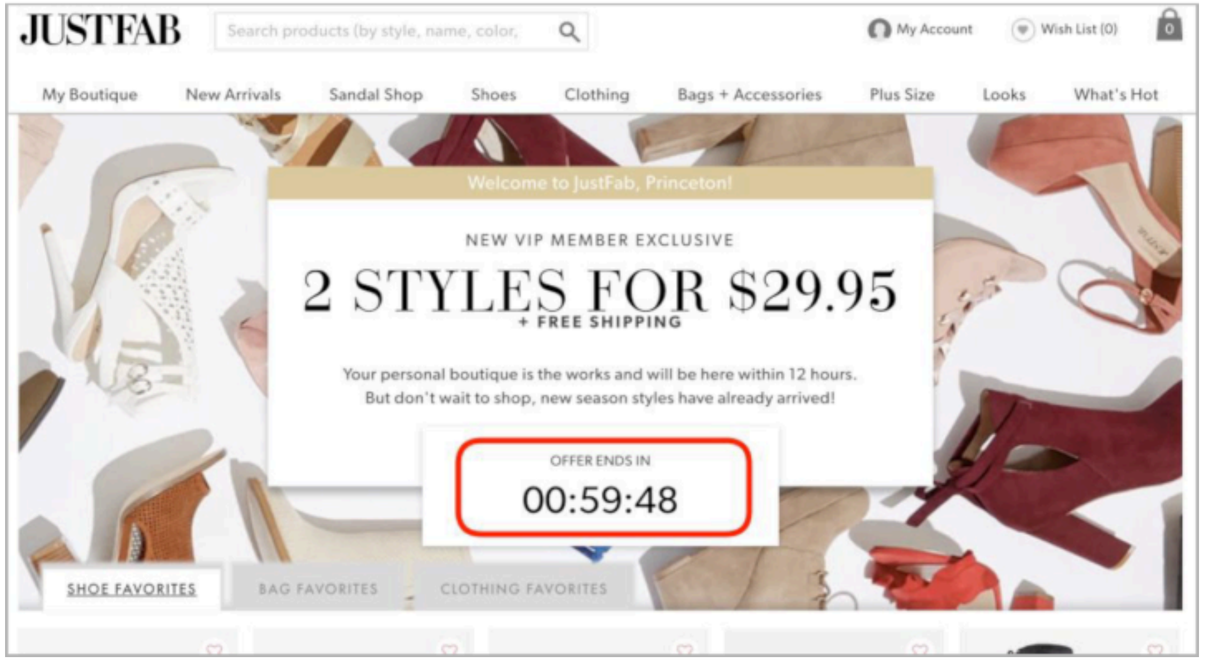



https://codecarbon.io/

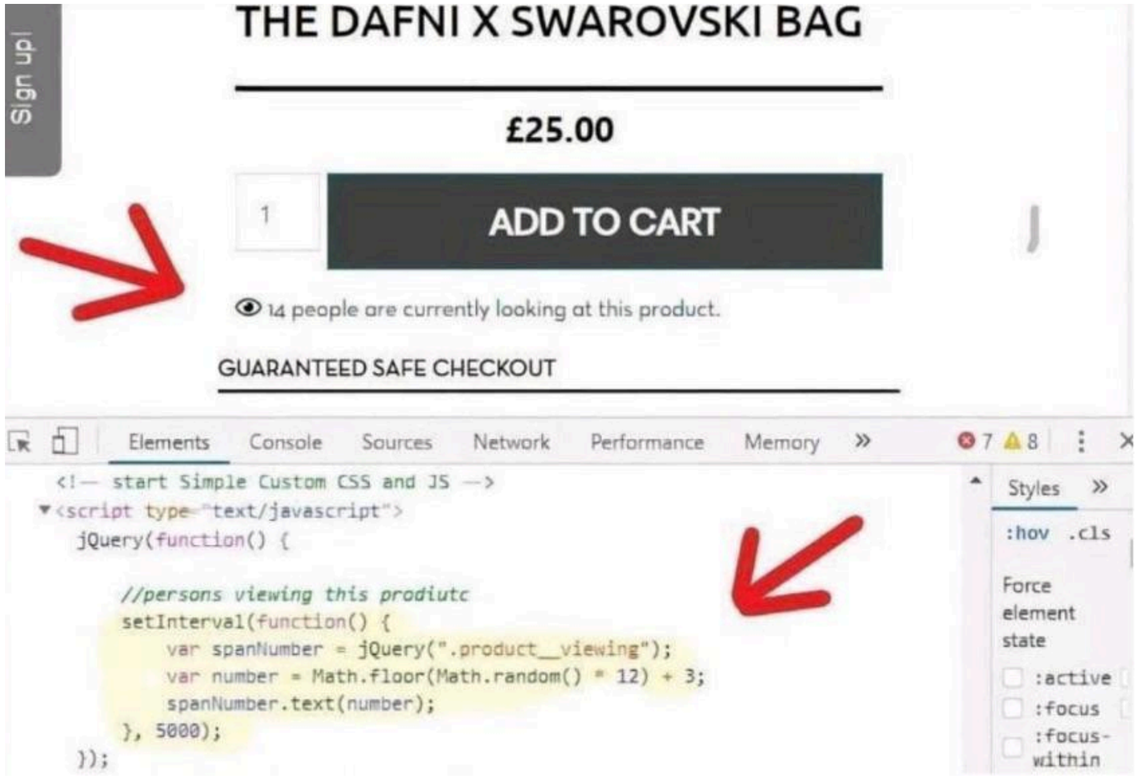
Learn to recognize and resist deceptive designs and dark patterns.

# Deception is tolerated in some areas of software.

FIGURE 1: A DECEPTIVE COUNTDOWN TIMER ON JUSTFAB




Narayanan et al., 2020




@random_walker

https://queue.acm.org/detail.cfm?id=3400901

# There's a slippery slope from inclarity to deception,

A (mostly) comprehensive history of Google's ad shading and labeling

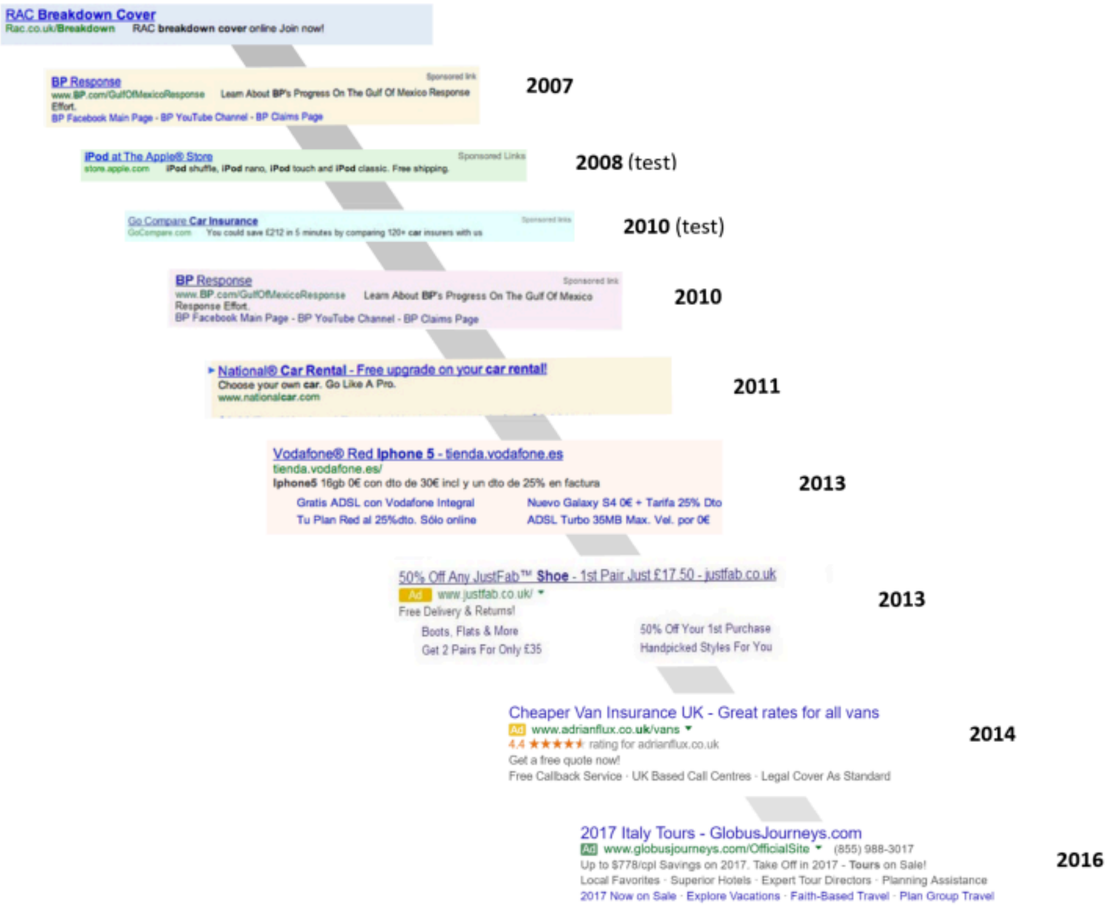
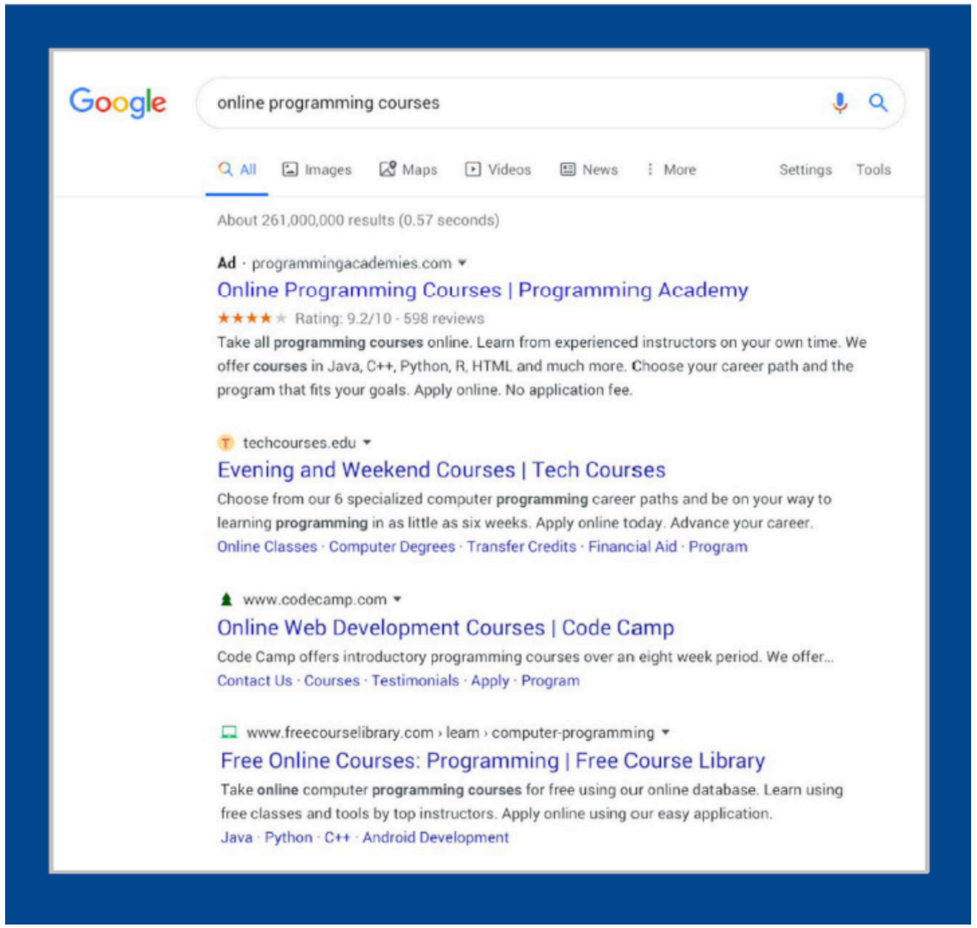



FIGURE 7: **GOOGLE'S RECENT CHANGE TO ITS AD LABELS**




2020

which allows for "frog-boiling".

# A number of "dark patterns" of deceptive design have emerged.

### Trick questions
While filling in a form you respond to a question that tricks you into giving an answer you didn't intend. When glanced upon quickly the question appears to ask one thing, but when read carefully it asks another thing entirely.

### Sneak into basket
You attempt to purchase something, but somewhere in the purchasing journey the site sneaks an additional item into your basket, often through the use of an opt-out radio button or checkbox on a prior page.

### Roach motel
You get into a situation very easily, but then you find it is hard to get out of it (e.g. a premium subscription).

### Privacy zuckering
You are tricked into publicly sharing more information about yourself than you really intended to. Named after Facebook CEO Mark Zuckerberg.

### Price comparison prevention
The retailer makes it hard for you to compare the price of an item with another item, so you cannot make an informed decision.

### Misdirection
The design purposefully focuses your attention on one thing in order to distract your attention from another.

### Hidden costs
You get to the last step of the checkout process, only to discover some unexpected charges have appeared, e.g. delivery charges, tax, etc.

### Bait and switch
You set out to do one thing, but a different, undesirable thing happens instead.

### Confirmshaming
The act of guilting the user into opting into something. The option to decline is worded in such a way as to shame the user into compliance.

### Disguised ads
Adverts that are disguised as other kinds of content or navigation, in order to get you to click on them.

### Forced continuity
When your free trial with a service comes to an end and your credit card silently starts getting charged without any warning. In some cases this is made even worse by making it difficult to cancel the membership.

### Friend spam
The product asks for your email or social media permissions under the pretence it will be used for a desirable outcome (e.g. finding friends), but then spams all your contacts in a message that claims to be from you.

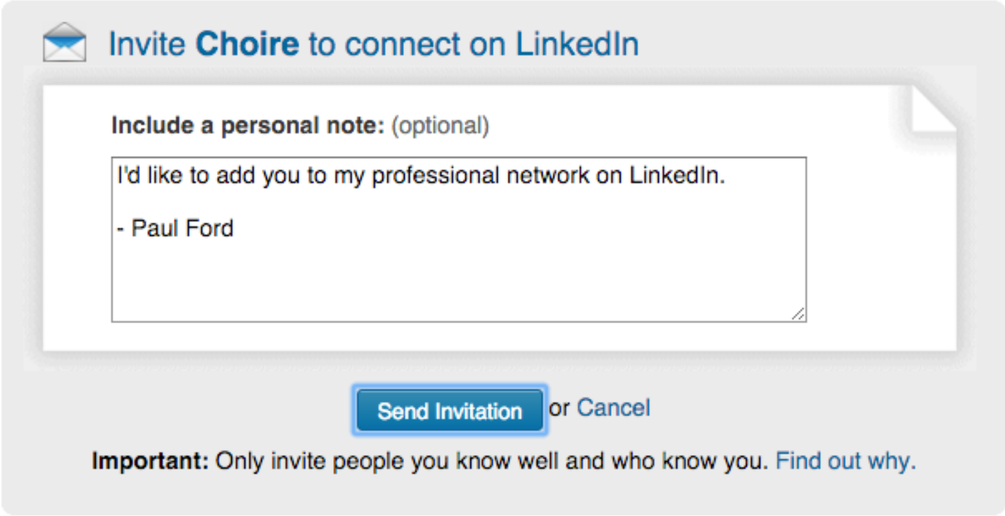
@darkpatterns

https://www.deceptive.design/

# "Growth hacking" is on shaky ethical and legal ground.

## LinkedIn pays big after class action lawsuit over user emails

By Adario Strange on October 3, 2015



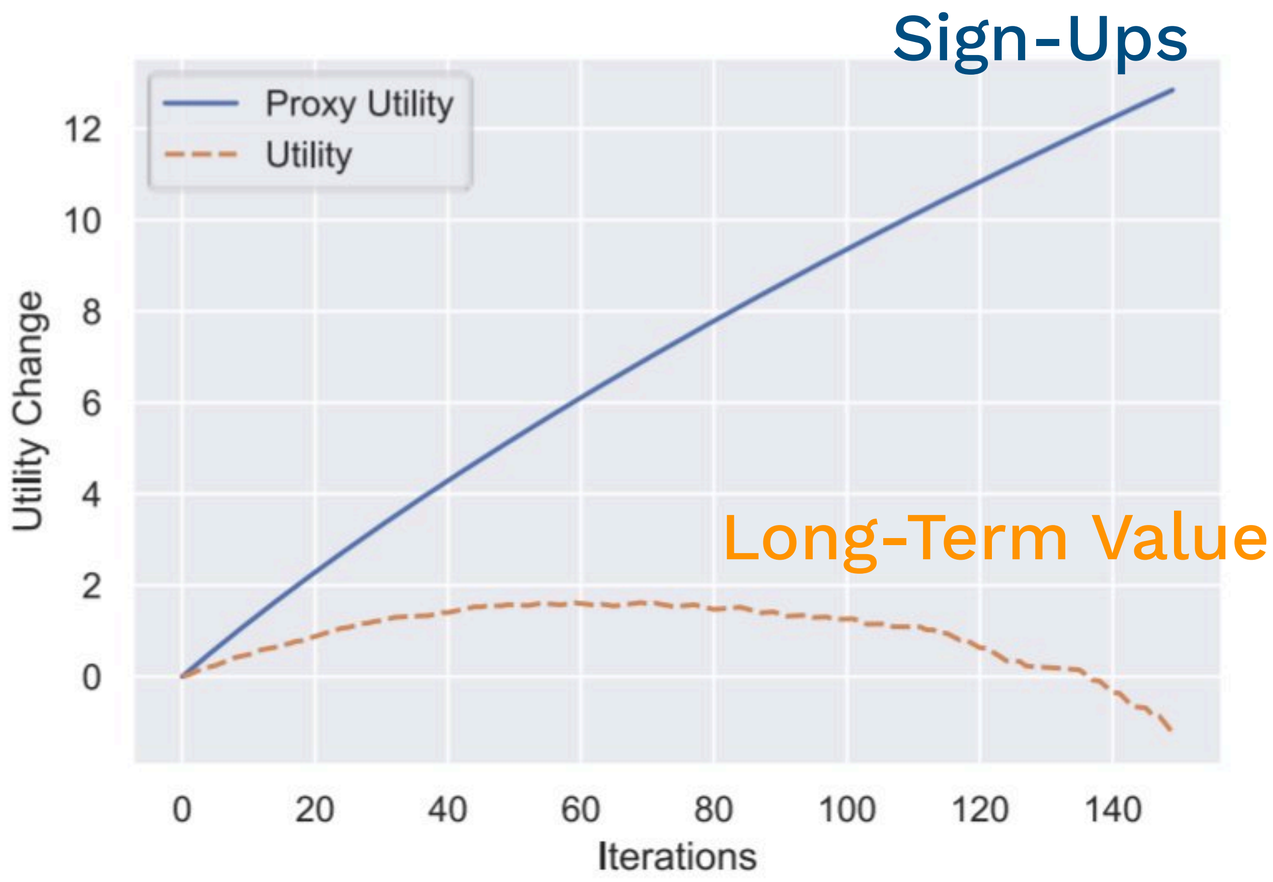

### The growth hack

Hotmail created a simple yet ingenious strategy to tackle that. They added a signature at the bottom of every email that'd go out from Hotmail. The signature was — **P.S. I love you. Get your free email at Hotmail.**

**Key Takeaway:** This is one of the best examples of turning your customers into your (unpaid) salespeople. The personal sounding nature of the message, along with the fact that it came from a friend and not the company itself, made clicking the link almost irresistible.

https://medium.com/@danrschlosser/linkedin-dark-patterns-3ae726fe1462

# ML can make this worse with A/B testing.




*Example A*

- [ ] I would like to receive the latest special offers, promotions and product information from AcmeCorp.
- [ ] I would like to receive carefully selected partner offers, competitions and messages.

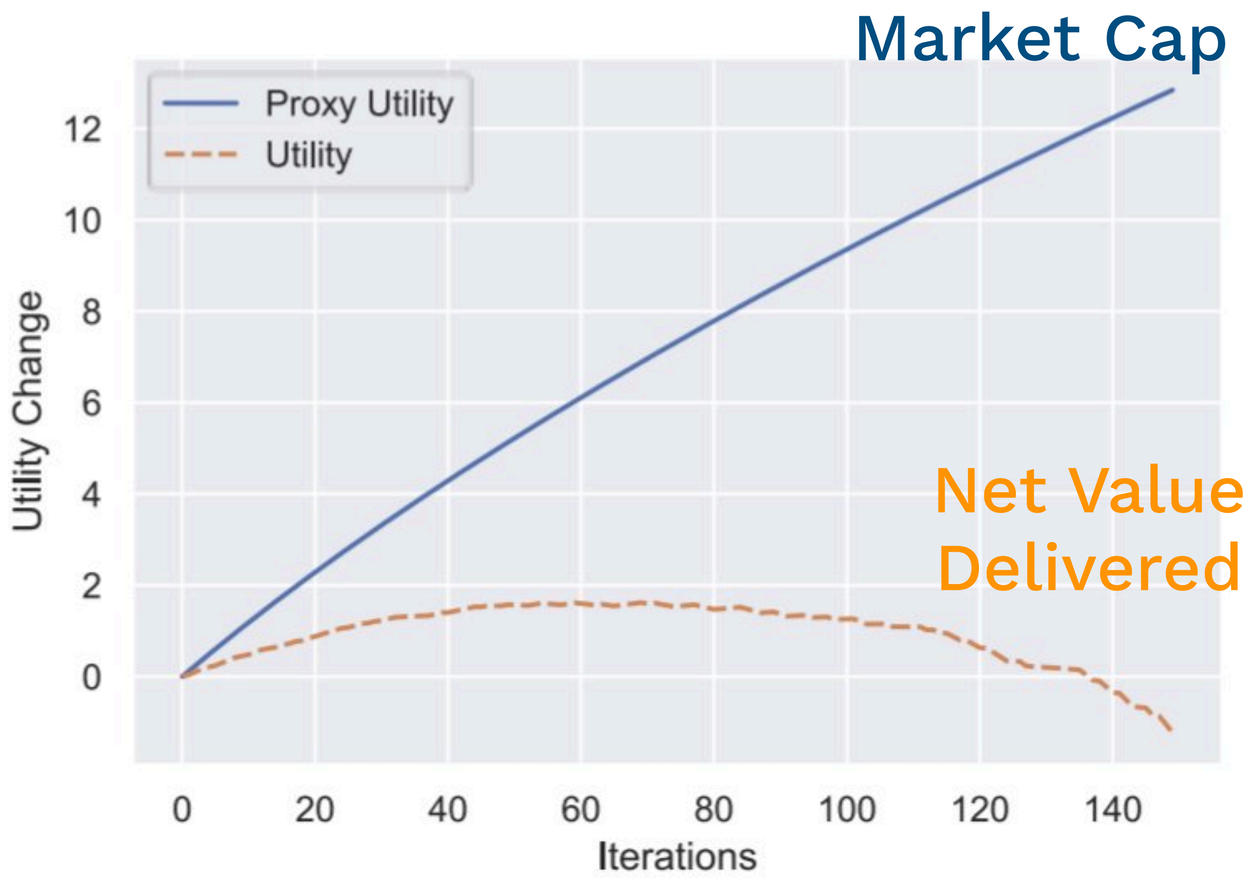
*Example B*

- [ ] Please tick here if you would prefer not to receive the latest special offers, promotions and product information from AcmeCorp
- [ ] I would like to receive carefully selected partner offers, competitions and messages.

## Alignment

https://alistapart.com/article/dark-patterns-deception-vs.-honesty-in-ui-design/

# These arise inside another alignment problem.




- One broadly-accepted justification for the private ownership of the means of production:

  private enterprise delivers broad social value, aligned by price signals and market forces.

- But private enterprises optimize metrics that are at best a proxy for social value.

- If you spend time at the intersection of funding, leadership, and tech, you will encounter this mismatch.

## 👯 Alignment

# What is to be done?

- Push for longer-term thinking within your org
  to allow for better alignment between metrics and goals
  and between goals and utility.

- Learn to recognize user-hostile designs
  and advocate for user-centered design instead.

# The tech industry cannot afford to ignore ethics.

- The tech industry should learn from other disciplines if it wants to avoid a trust crisis.
- We can start with educating ourselves about common deceptive or user-hostile practices in our industry.

02

# ML Ethics

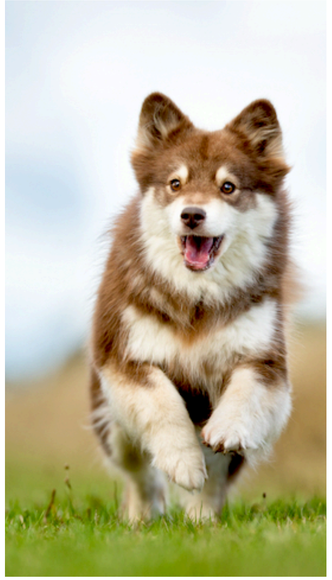
# Ethical questions are coming to the fore in machine learning.

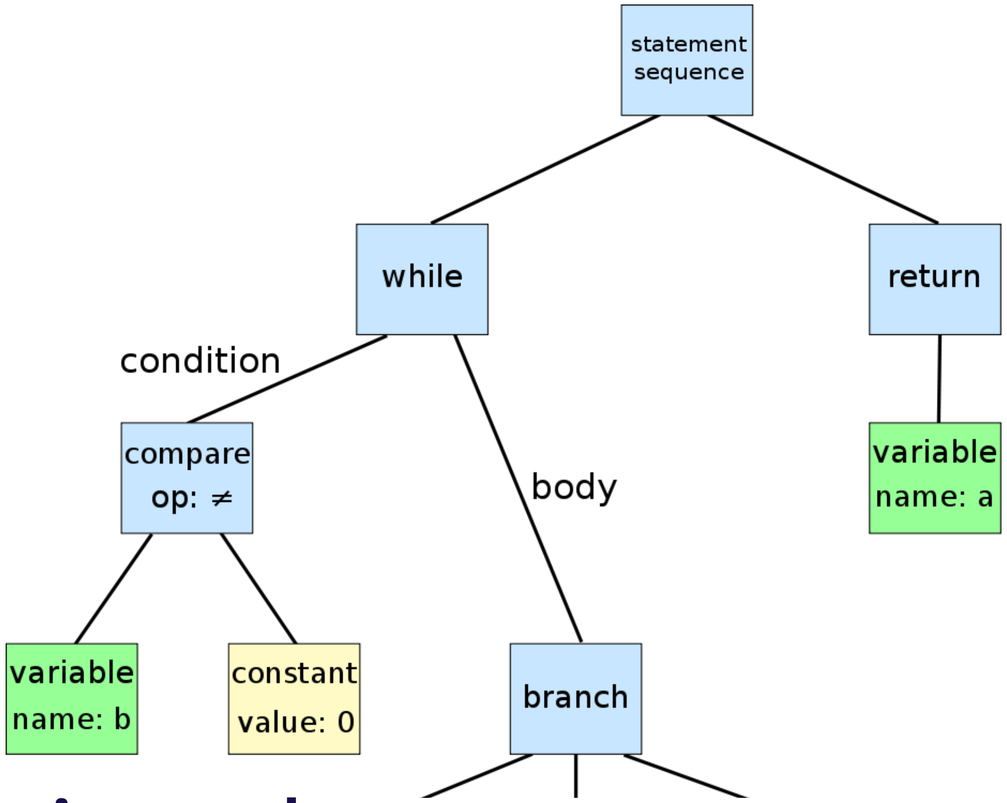
- Why not just tech ethics?
- Common ethical questions in machine learning
- What we can learn from Medical ML

Machine learning touches human lives more intimately.

# Many ML methods make human-legible data into computer-legible data.

- Computer *vision*
- *Natural* language processing






Humans are more sensitive to errors in and have more opinions about *images* than they do *ASTs*.

## ⚖️ Trade-Offs

# ML involves being wrong all the time.

- Our models are statistical and include "randomness".

- Randomness is almost always an admission of ignorance.
  - Applies even to quintessential examples of "randomness", like RNG and coin flips.

- It is upsetting and even harmful to be misunderstood by ML.

Certain ethical questions come up frequently and have established camps.

# You should have thoughtful answers to these questions.

- Is the model "fair"?
- Is the system accountable?
- Who owns the data?
- Should this be built at all?

# Is the model "fair"?

# The classic case comes from criminal justice.

**COMPAS**

Correctional Offender Management Profiling for Alternative Sanctions

- Goal: before trial, predict whether defendants will be arrested again.
  - Operationalization: 10 point re-arrest probability based on past data.

- Sub-goal: be *less biased* than human judges.
  - Operationalization: arrest probabilities are calibrated for subgroups.

# An independent audit revealed a concerning pattern.

## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica*
May 23, 2016

### Prediction Fails Differently for Black Defendants

| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

*Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)*

More false positives for Black defendants,
more false negatives for White defendants.

Humility

# Some quick algebra reveals bias is inevitable.

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR})$$

p: Probability of re-arrest
FP: False Positive, "extra imprisonment"
FN: False Negative, "missed opportunity"
PP: Positive Prediction, "avoided re-arrest"

Chouldechova, 2017

https://arxiv.org/abs/1610.07524

# Some quick algebra reveals bias is inevitable.

$$R_1 = \frac{A_1}{B_1 \cdot C_1} \qquad R_2 = \frac{A_2}{B_2 \cdot C_2}$$

R: Re-arrest ratio, p / 1-p
A: False positive rate, FPR
B: (1 - PPV) / PPV
C: True positive rate, 1 - FNR

Chouldechova, 2017

https://arxiv.org/abs/1610.07524

# See "21 Definitions of Fairness" for more.

Translation tutorial:

21 fairness definitions and their politics

Arvind Narayanan

@random_walker

## ⚖️ Trade-Offs

https://www.youtube.com/watch
?v=jIXIuYdnyyk&ab_channel=ArvindNarayanan

@random_walker

# If prevalences differ across groups, one of FPR, FNR, and PPV is not equal. Unfair!

$$\text{FPR} = \frac{p}{1-p}\frac{1-\text{PPV}}{\text{PPV}}(1-\text{FNR})$$

larger for Black defendants

equalized across groups in COMPAS

**Prediction Fails Differently for Black Defendants**

| | WHITE | AFRICAN AMERICAN |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

*Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)*

Chouldechova, 2017

https://arxiv.org/abs/1610.07524

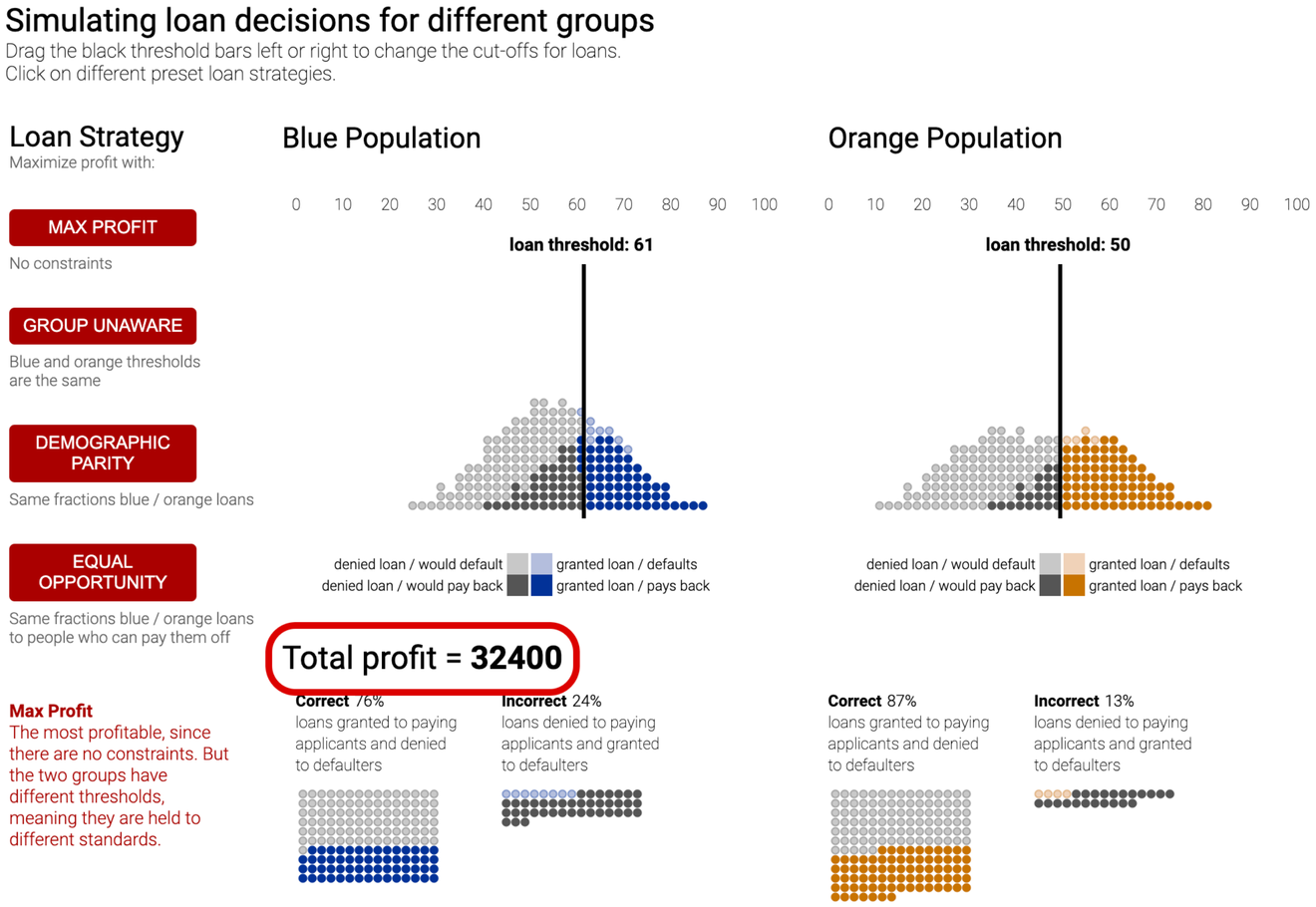
# But the impact of "unfairness" is not fixed.

- This story is often presented as:
  "no matter what, they would've found something to complain about".

- But note: equalizing FPR and PPV across groups would lead to
  a higher FNR for Black defendants relative to White defendants.

- In the context of American politics, that's not going to lead to complaints
  from the same people.

# Try making these trade-offs for yourself!




## ⚖️ Trade-Offs

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

# This kicked off a flurry of research on fairness.

FAccT 서울 2022

Fairness,
ACCountability, &
Transparency

# But should this model be built at all?

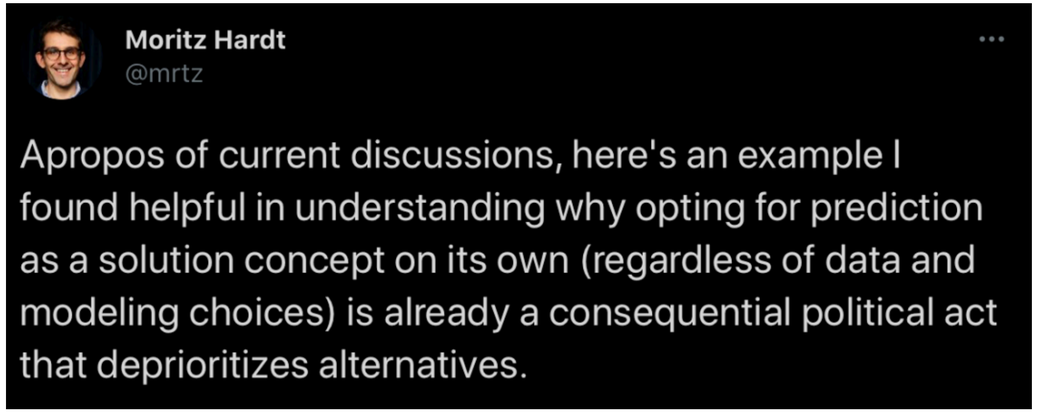
COMPAS is proprietary and uninterpretable.
It has an accuracy of ~65%.

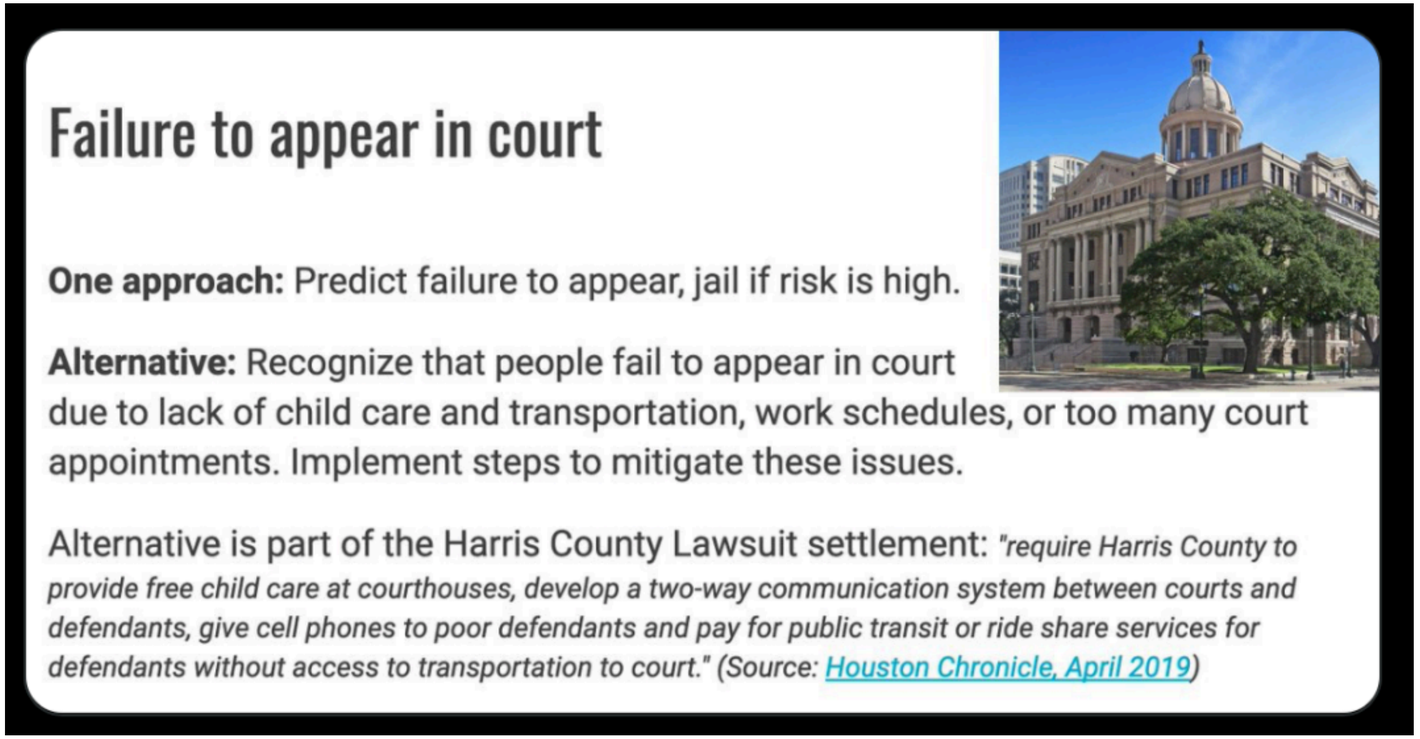
Pooling non-experts (like a jury) has an accuracy of ~65%.

Interpretable "scores" based on age and prior arrests
have an accuracy of ~65%.

https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/

# But should this model be built at all?



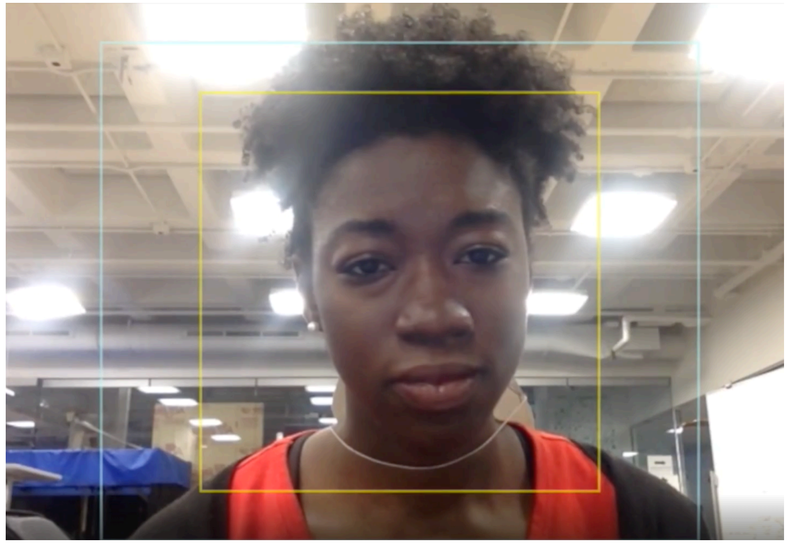



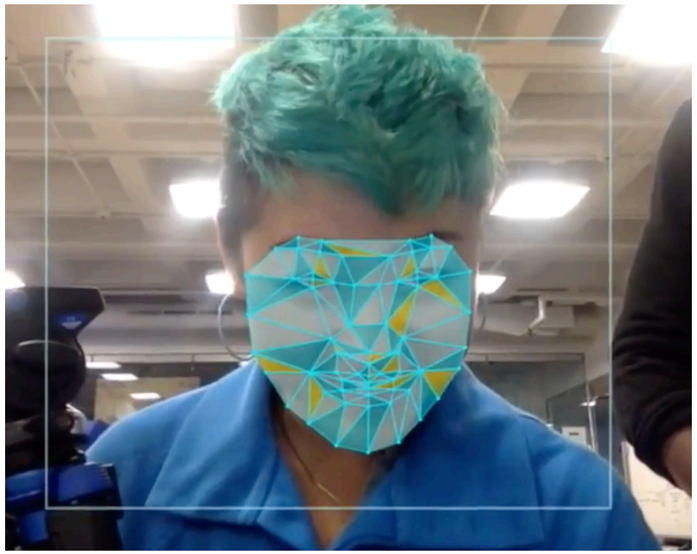

Re-arrest is not recidivism!

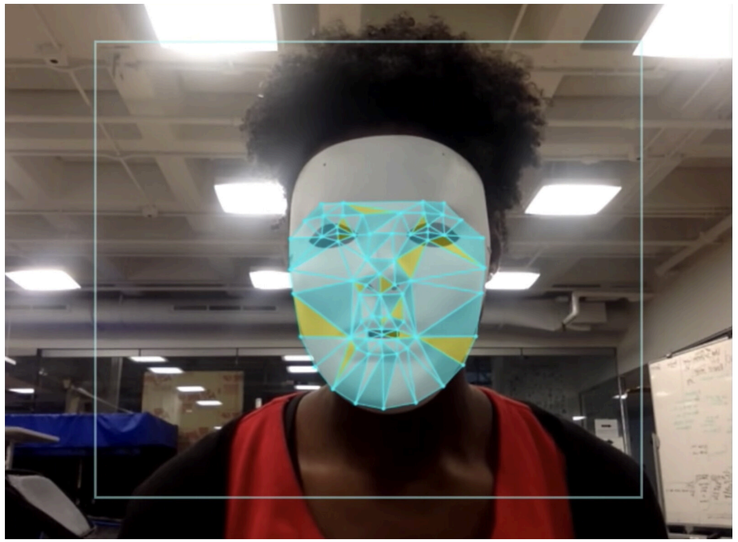
Alignment

# Representation matters for model fairness.

# It is easy to make ML-powered tech that fails for minoritized groups.



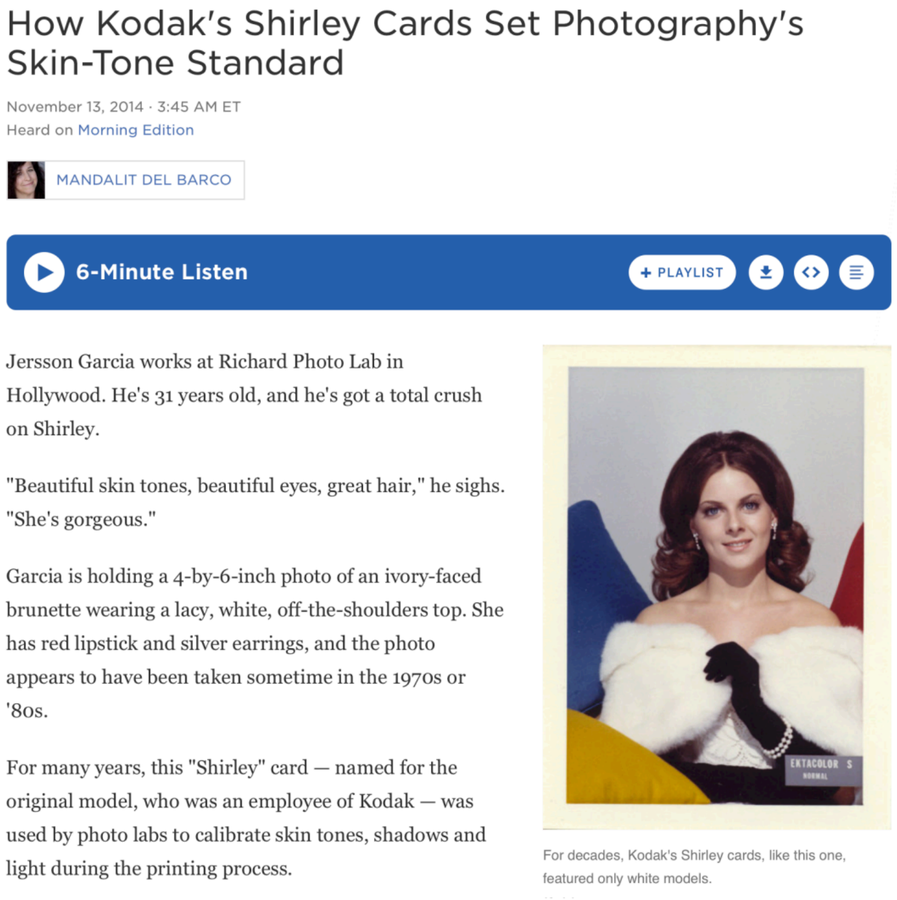




## Off-the-shelf computer vision failed on darker skin.

https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

# This is not a new issue, just more salient.




https://twitter.com/nke_ise/status/897756900753891328

## ⚖️ Trade-Offs

# Consider model and dataset cards.

## Face Detection

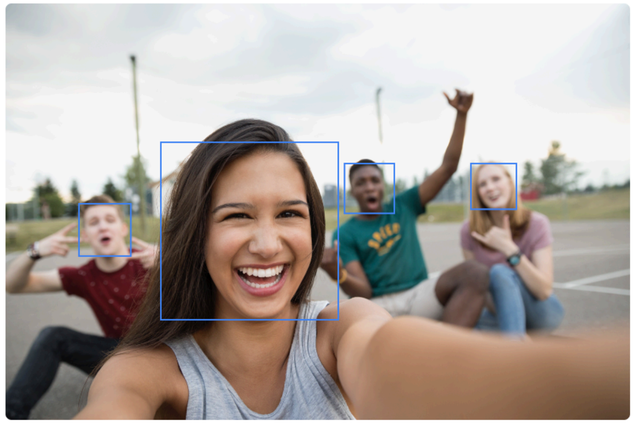
Model Card v0 Cloud Vision API

- Overview
- Limitations
- Trade-offs
- Performance
- Test your own images
- Provide feedback

**Explore**

- Object Detection

MODEL DESCRIPTION



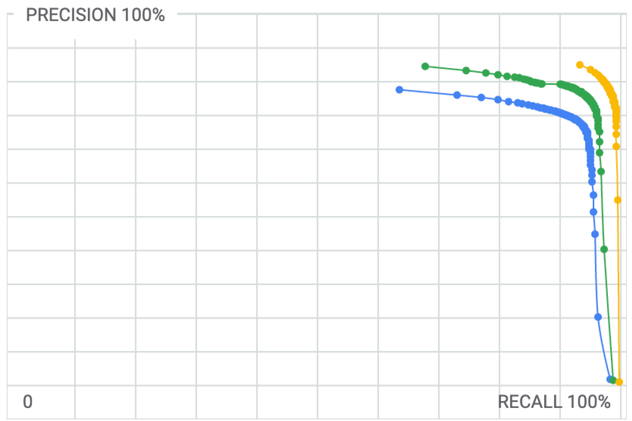
**Input**: Photo(s) or video(s)

**Output**: For each face detected in a photo or video, the model outputs:

- Bounding box coordinates
- Facial landmarks (up to 34 per face)
- Facial orientation (roll, pan, and tilt angles)
- Detection and landmarking confidence scores.

No identity or demographic information is detected.

PERFORMANCE




Open Images · Face Detection Dataset Benchmark · Labeled Faces in the Wild

Overall model performance, and performance sliced by different image and face characteristics, were assessed, including:

- Derived characteristics (face size, facial orientation, and occlusion)
- Face demographics (human-perceived gender presentation, age, and skin tone)

# Automatically generated in Hugging Face.

https://modelcards.withgoogle.com/about @timnitGebru

# Diversify your networks.

@black_in_ai

WOMEN IN DATA SCIENCE

https://www.widsconference.org/

@WiDS_Worldwide

DIVERSIFY TECH

**Job Board**
**for underrepresented folks in tech**

https://www.diversifytech.co/job-board/

@DiversifyTechCo

https://anchor.fm/ucla-acm-ai

# Solving these problems is not easy.

A spokesperson for OpenAI told *New Scientist* that prompts given to DALL-E 2 were modified if they were "underspecified". If a prompt describes a generic person and doesn't specify what gender or race they should be, then DALL-E 2 will be specifically told to add a certain race and gender "with weights based on the world's population", said the spokesperson. The company declined to grant access to DALL-E 2 so that *New Scientist* could run its own tests.

👯 Alignment

🙇‍♀️ Humility

https://www.newscientist.com/article/2329690

# Solving these problems is not easy.

## World Faces

**What would a hundred people randomly selected from around the world look like?**

The images were generated using **StableDiffusion on Replicate** with prompts like *"profile headshot photo of a male latin american adult, centered, white bg"*.

By **Sergey Karayev**. Code on **GitHub**.

https://www.worldfaces.xyz/

Humility

Only some people get headshots taken.

Who does this make invisible?

# Is the system accountable?

# Some consider explanation and accountability in the face of important judgments to be human rights.

the right … **to obtain an explanation** of the decision reached after such assessment and **to challenge the decision**.

Recital 71

(legal status is currently unclear, being litigated)

https://arxiv.org/abs/1606.08813

# This is not conceptually new or even tech-specific.

## Equal Credit Opportunity Act of 1974

Paragraph 9(b)(2).

1. **Number of specific reasons.** A creditor must disclose the principal reasons for denying an application or taking other adverse action. The regulation does not mandate that a specific number of reasons be disclosed, but disclosure of more than four reasons is not likely to be helpful to the applicant.

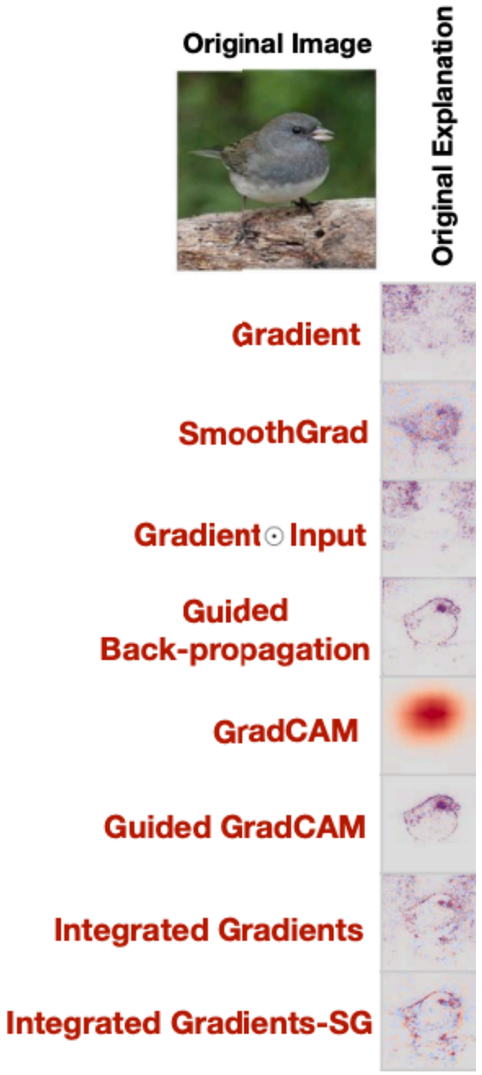
2. **Source of specific reasons.** The specific reasons disclosed under §§ 1002.9(a)(2) and (b)(2) must relate to and accurately describe the factors actually considered or scored by a creditor.

3. **Description of reasons.** A creditor need not describe how or why a factor adversely affected an applicant. For example, the notice may say "length of residence" rather than "too short a period of residence."

https://www.consumerfinance.gov/rules-policy/regulations/1002/interp-9/#9-b-1-Interp-1

# There are "off-the-shelf" methods for introspecting DNNs.






The input-output gradient captures *local* contributions.

**Add noise to input, average the results**

Multiply by input magnitude

Hack in a different gradient for ReLU

**Classification only: average grads of last layer**

GradCAM + Hack in a different gradient for ReLU

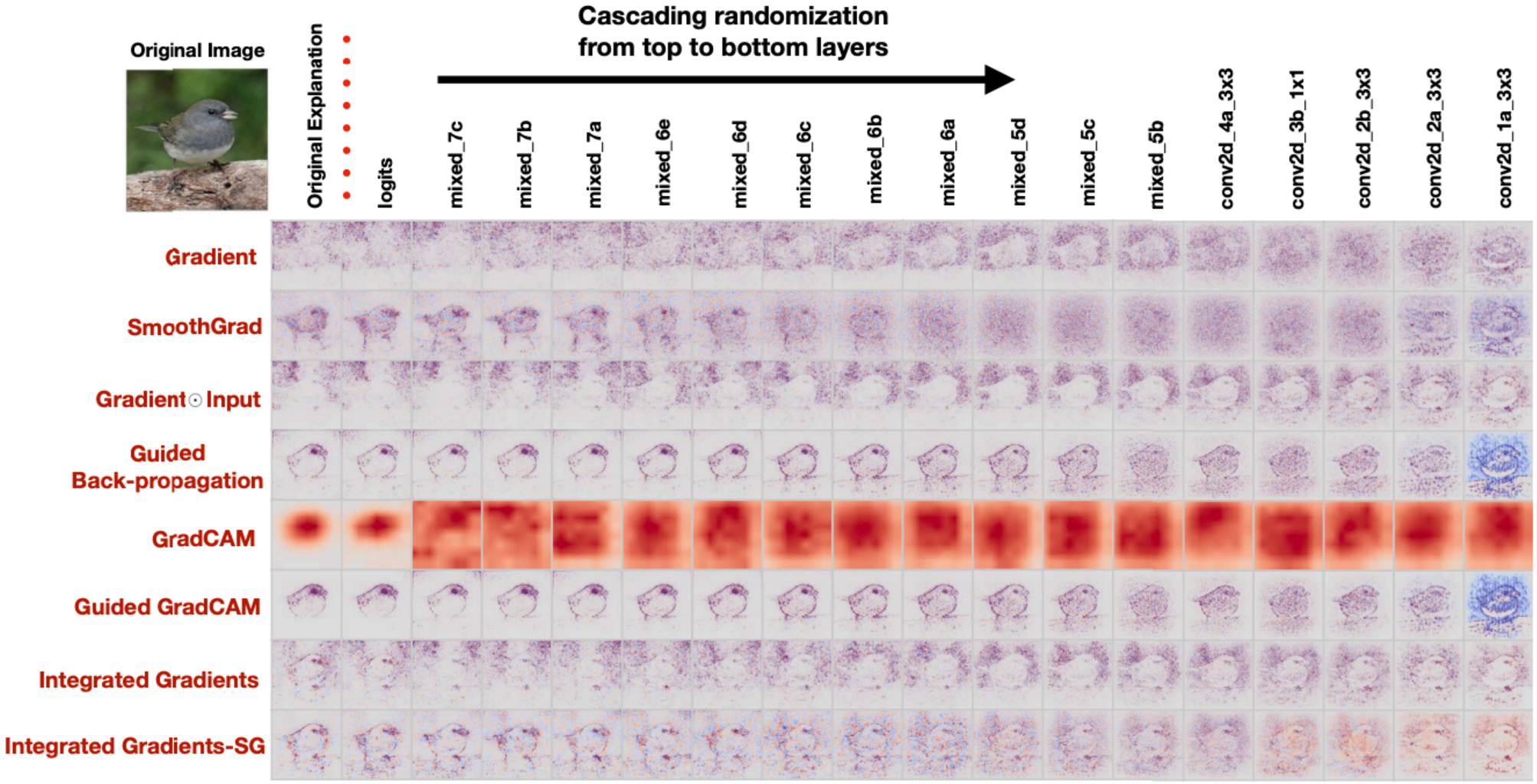
**Integrate the gradients along a path from baseline.**

Integrated Gradients with noise

Adebayo et al., 2018

https://arxiv.org/abs/1810.03292

# Bad news: they aren't very robust.




Adebayo et al., 2018

https://arxiv.org/abs/1810.03292

# Introspecting DNNs requires "reverse engineering".

Distill

ABOUT PRIZE SUBMIT

March 4, 2021

PEER-REVIEWED

## Multimodal Neurons in Artificial Neural Networks

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah

We report the existence of multimodal neurons in artificial neural networks, similar to those found in the human brain.

Nov. 17, 2020

PEER-REVIEWED

## Understanding RL Vision

Jacob Hilton, Nick Cammarata, Shan Carter, Gabriel Goh, and Chris Olah

With diverse environments, we can analyze, diagnose and edit deep reinforcement learning models using attribution.

https://distil.pub

<EOT>Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs Potter was Mrs Dursley's sister, but they hadn't met for several years; in fact, Mrs Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbours would say if the Potters arrived in the street. The Dursleys knew that the Potters had a small son, too, but they had never even seen him. This boy was another good reason for keeping the Potters away; they didn't want Dudley mixing with a child like that.

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr Dursley was the director of a firm called Grunnings, which made drills.

MARCH 2022

### *In-Context Learning and Induction Heads*

*Olsson et al., 2022*

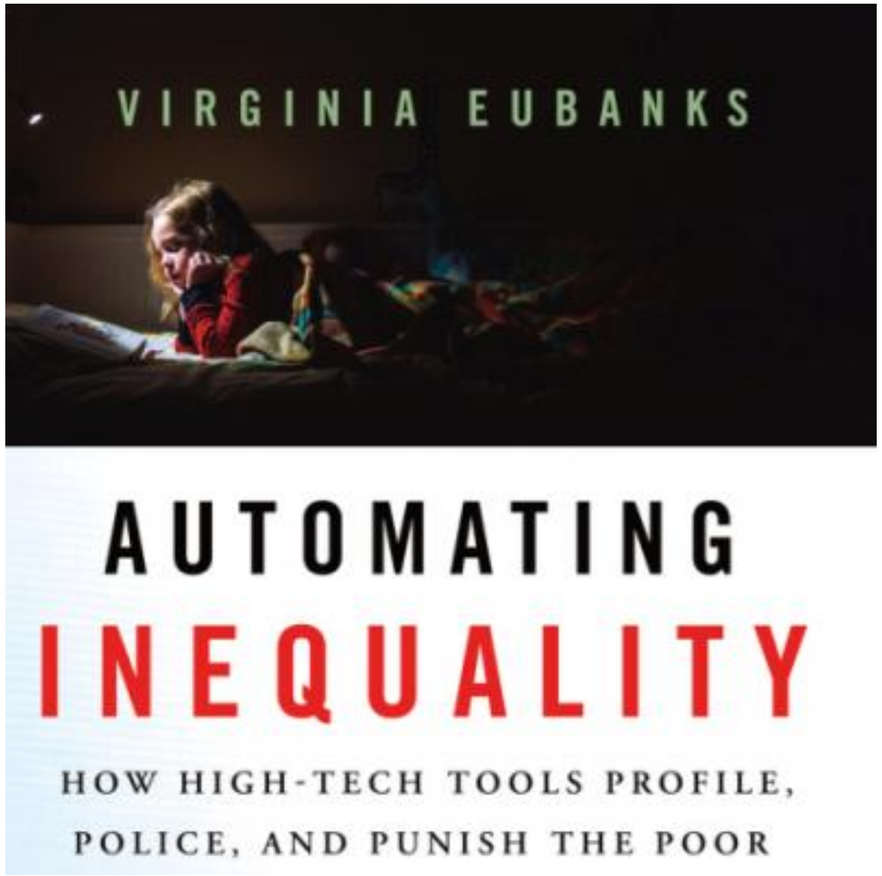
An exploration of the hypothesis that induction heads are the primary mechanism behind in-context learning. We also report the existence of a previously unknown phase change in transformers language models.

https://transformer-circuits.pub/

@ch402

# This is still a research problem.

# Technical systems are prone to unaccountability.




## ⚖️ Trade-Offs

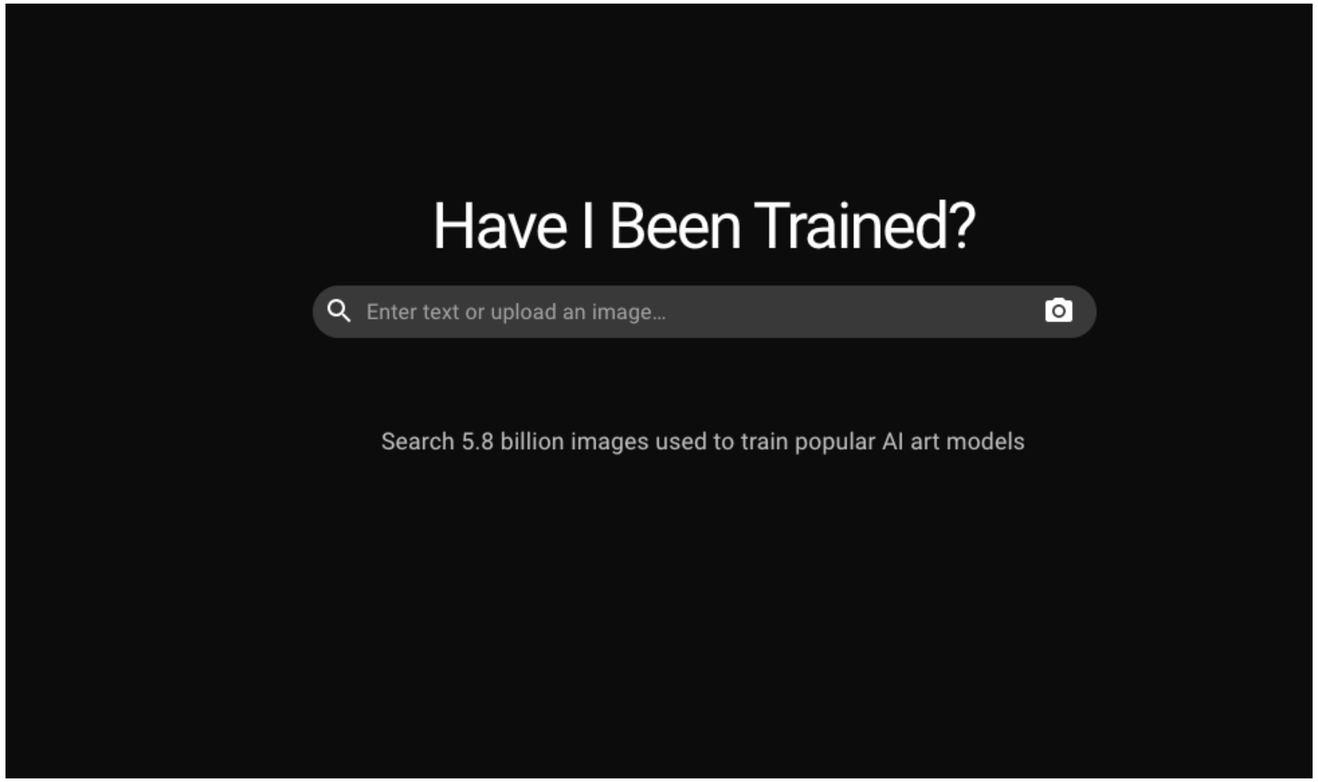
So ask: should this be built at all?
Who benefits and who is harmed by automation?

# Who owns data?

# Training data is generated by humans, who generally feel some ownership over that data.

- Large datasets are often produced by "crawling" the internet
- Many people are not aware that this is both possible and legal
- As technology has changed, what can be done with data has changed

# Image-to-Text models are trained on illegal data.




https://haveibeentrained.com/

https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/

## ⚖️ Trade-Offs

**RJ Palmer ➡ Lightbox Expo #510**
@arvalis

What makes this AI different is that it's explicitly trained on current working artists. You can see below that the AI generated image(left) even tried to recreate the artist's logo of the artist it ripped off.

This thing wants our jobs, its actively anti-artist.

25.4K 6:16 PM - Aug 13, 2022

**Logan Preshaw**
@wickedinsignia

Those claiming that artists are "gatekeeping art" don't understand that the developers behind these tools never had the right to the data that defines their software. Artists are solely responsible for what these machines can do, and are solely displaced by them.

2,086 9:44 PM - Sep 11, 2022

**Jesse Engel**
@jesseengel

This was one of my biggest challenges with Jukebox as well. Training, but especially conditioning on individual artists without their consent feels like it crosses a line for me personally. It's ethically a grey area at the least, and advertising the ability feels exploitative. twitter.com/arvalis/status…

# And "technically you agreed to this" doesn't satisfy people.

## Recall the Facebook emotion research study.

# This is likely to be a new frontier of data governance.

He said the project was working with "technology industry leaders to create mechanisms by which artists can upload their portfolios and request for their styles not to be used in online services using this and similar technology".

Emad Mostaque of Stable Diffusion

**Holly Herndon**
@hollyherndon

Some people seem to be confused by that slide outside of its context in the presentation, so I updated it.

It is meant to establish significant differences in how sampling and spawning work, using contemporary AI systems

Differences between Sampling and Spawning (AI) systems

| 20th Century Samplers | 21st Century Spawners (AI Systems) |
|---|---|
| Individuals sample others with no means to hard code attribution | Spawning uses training data of others that can be attributed to the source |
| Individuals sample others with no means to hard code payment | Spawners could pay for training data use at source, or share profit when profit is made |
| Sampling recontextualizes fixed sounds within new compositions | Spawning generates entirely new sounds or compositions from training material. |
| Sampling is one to one reproduction of a recorded sound, allowing minor alterations | Spawning can generate infinite sounds from analysis of training material, it is reproductive |

26 7:57 AM - Sep 5, 2022

## ⚖️ Trade-Offs

# Dataset cards are helpful here!

**Dataset Structure**

- Data Instances
- Data Fields
- Data Splits

**Dataset Creation**

- Curation Rationale
- Source Data
- Annotations
- Personal and Sensitive Informa…

## Dataset Card for CNN Dailymail Dataset

### Dataset Summary

The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine reading and comprehension and abstractive question answering.

https://huggingface.co/docs/datasets/dataset_card @timnitGebru

# Check out the deon "ethics checklist".

deon

*Deon adds an ethics checklist to your data science projects.*

About

Examples

- Where things have gone wrong

## Where things have gone wrong

To make the ideas contained in the checklist more concrete, we've compiled examples of times when things have gone wrong. They're paired with the checklist questions to help illuminate where in the process ethics discussions may have helped provide a course correction.

| Checklist Question | Examples of Ethical Issues |
| --- | --- |

https://deon.drivendata.org/examples/

They have also collected some nice failure cases.

# Should this be built at all?

# ML-powered weaponry is already here.

A remote-controlled weapon that uses CV for targeting. Hebron, West Bank.

“In its pilot stage” the system will use only sponge-tipped bullets.

Note that these can still cause serious injury.

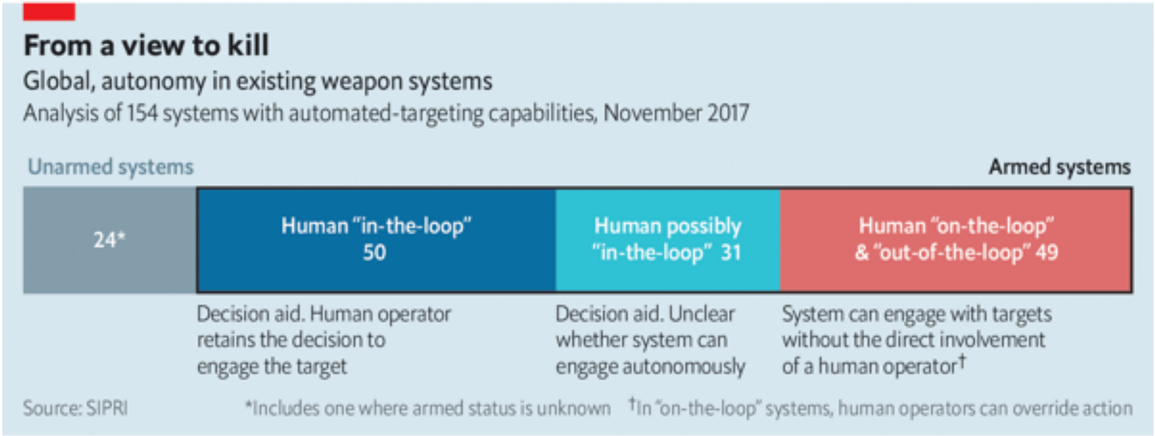
# The definition of "autonomous weapon" is blurry.

The *Harop* drone is a "loitering munition" that can automatically destroy radar that locks onto it.

It was used in the Nagorno-Karabakh War in 2021.

The Phalanx CIWS automatically fires at targets moving towards naval vessels at very high velocities.

Used in the First Gulf War in 1991.



# Perhaps they are nearly 100 years old.

German "Bouncing Betty" mine

Anti-personnel mines saw widespread use starting in the 1930s.

**Ottawa Treaty**
**(Mine Ban Treaty)**

## Or older, if you count traps.

# But that doesn't mean they are ethical.

**Ottawa Treaty**
**(Mine Ban Treaty)**

**Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction**

States that are party to the Ottawa Treaty

CAMPAIGN TO STOP KILLER ROBOTS

Perhaps rather than building and then banning, we could just not build.

# You should have thoughtful answers to these questions.

- Is the model "fair"?
  - Fairness is possible, but requires trade-offs.
- Is the system accountable?
  - Accountability is easier than interpretability.
- Who owns the data?
  - Answer this up front. Changes are on the way.
- Should this be built at all?
  - Repeatedly ask this.
  - Use this to narrow scope.

The broader ML world
can learn a lot from medicine.

# The mismatch between ML and medicine has led to much learning.

- COVID-19 and ML: a fiasco
- Why medicine?
- Auditing and frameworks

Medical researchers found that
almost all ML research on COVID-19 was useless.

# The biomedical response to COVID-19 was a triumph.

## Global impact of the first year of COVID-19 vaccination: a mathematical modelling study

Oliver J Watson, PhD † • Gregory Barnsley, MSc † • Jaspreet Toor, PhD • Alexandra B Hogan, PhD • Peter Winskill, PhD • Prof Azra C Ghani, PhD • Show footnotes

Open Access • Published: June 23, 2022 • DOI: https://doi.org/10.1016/S1473-3099(22)00320-6 •

Check for updates

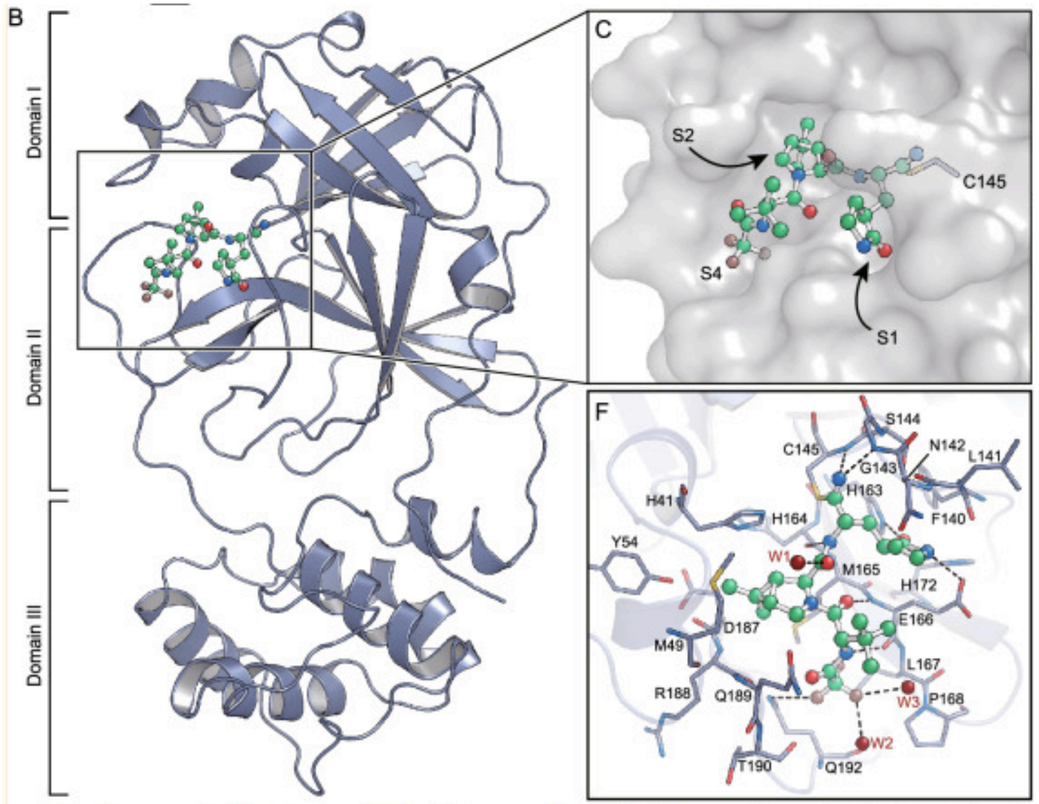
"Based on official reported COVID-19 deaths,
we estimated that vaccinations prevented 14.4 million deaths"

Article | Published: 05 August 2020

## SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness

Kizzmekia S. Corbett, Darin K. Edwards, Sarah R. Leist, Olubukola M. Abiona, Seyhan Boyoglu-Barnum, Rebecca A. Gillespie, Sunny Himansu, Alexandra Schäfer, Cynthia T. Ziwawo, Anthony T. DiPiazza, Kenneth H. Dinnon, Sayda M. Elbashir, Christine A. Shaw, Angela Woods, Ethan J. Fritch, David R. Martinez, Kevin W. Bock, Mahnaz Minai, Bianca M. Nagata, Geoffrey B. Hutchinson, Kai Wu, Carole Henry, Kapil Bahl, Dario Garcia-Dominguez, ... Barney S. Graham + Show authors

*Nature* **586**, 567–571 (2020) | Cite this article




Crystal structure of SARS-CoV-2 main protease in complex with [Paxlovid]. Zhao et al., 2022.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8533666/

# At a crucial time, ML did not acquit itself well.

Roberts et al., 2021

## Conclusions

This systematic review specifically considers the current machine learning literature using CT and CXR imaging for COVID-19 diagnosis and prognosis, which emphasizes the quality of the methodologies applied and the reproducibility of the methods. We found that no papers in the literature currently have all of: (1) a sufficiently documented manuscript describing a reproducible method; (2) a method that follows best practice for developing a machine learning model; and (3) sufficient external validation to justify the wider applicability of the

https://www.nature.com/articles/s42256-021-00307-0

**Conclusion** Prediction models for covid-19 entered the academic literature to support medical decision making at unprecedented speed and in large numbers. Most published prediction model studies were poorly reported and at high risk of bias such that their reported predictive performances are probably optimistic. Models with low risk of bias should be validated before clinical implementation,

Wynants et al., 2020-2022 https://www.bmj.com/content/369/bmj.m1328

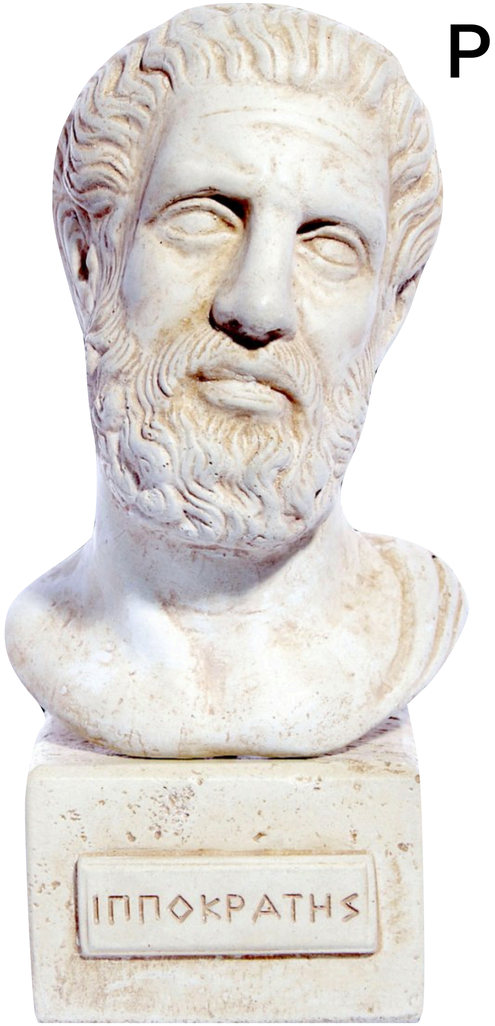
Humility

Source issues. Many papers (16/62) used the pneumonia dataset of Kermany et al.[86] as a control group. They commonly failed to mention that this consists of paediatric patients aged between one and five. Developing a model using adult patients with COVID-19 and very young patients with pneumonia is likely to overperform as it is merely detecting children versus adults. This

**Roberts et al., 2021**

https://www.nature.com/articles/s42256-021-00307-0

Medicine has a strong culture of ethics
that equips it to handle very serious problems.




Practice two things in your dealings with disease: either help or do not harm the patient.

Move fast and break things.

# When this culture is applied to medicine, results can be ugly.

**Jonathan Zittrain**
@zittrain

This @newsbeagle and @meharris piece describes a retinal implant used by hundreds of people to gain sight -- one that has been declared by its vendor to be obsolete, leaving customers -- patients -- with no recourse against hardware or software bugs or issues. h/t @Williamjmarks5

> **BARBARA CAMPBELL** was walking through a New York City subway station during rush hour when her world abruptly went dark. For four years, Campbell had been using a high-tech implant in her left eye that gave her a crude kind of bionic vision, partially compensating for the genetic disease that had rendered her completely blind in her 30s. "I remember exactly where I was: I was switching from the 6 train to the F train," Campbell tells *IEEE Spectrum*. "I was about to go down the stairs, and all of a sudden I heard a little 'beep, beep, beep' sound."
>
> It wasn't her phone battery running out. It was her Argus II retinal implant system powering down. The patches of light and dark that she'd been able to see with the implant's help vanished.

260 7:44 AM - Aug 28, 2022

**Arvind Narayanan**
@random_walker

When tech culture comes for things that actually matter, like medical devices, you get horrors like blind people's $150,000 bionic implants going kaput because the company stopped supporting the software on them, leaving them blind again. This issue deserves a lot more attention.
twitter.com/klaudiajaz/sta…

> **klaudia jaźwińska** @klaudiajaz
> In my new op-ed in @statnews, I make the case for why regulators need to take swift action to protect users from being cut off from life-saving or life-changing technologies as a result of business failure.
> statnews.com/2022/08/10/imp…

70 10:32 AM - Aug 10, 2022

21 people are talking about this

https://www.statnews.com/2022/08/10/implant-recipients-shouldnt-be-left-in-the-dark-when-device-company-moves-on/

Researchers at the intersection of medicine and ML are proposing solutions.

# Clinical trial standards have been extended to ML.

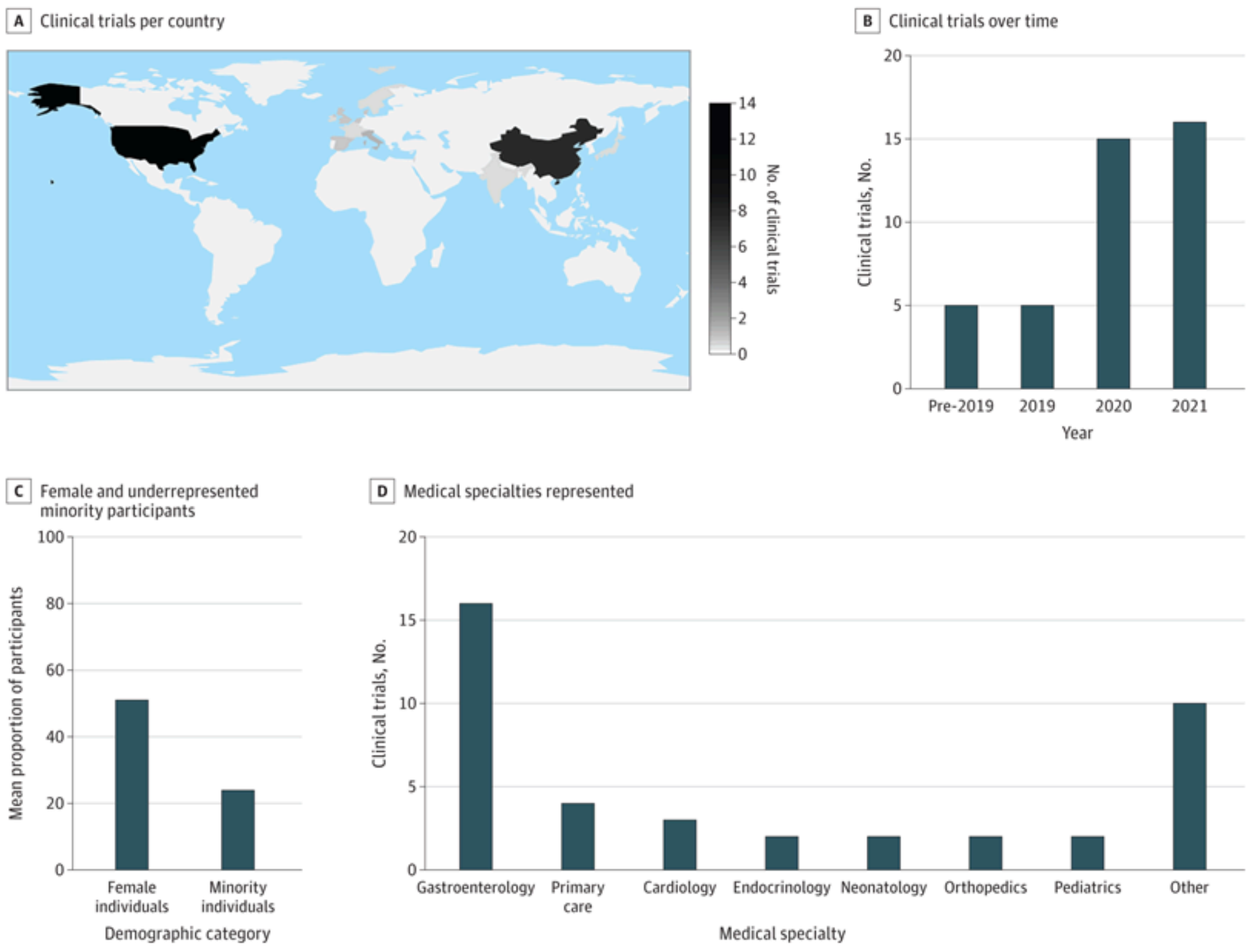
SPIRIT: Protocol design.

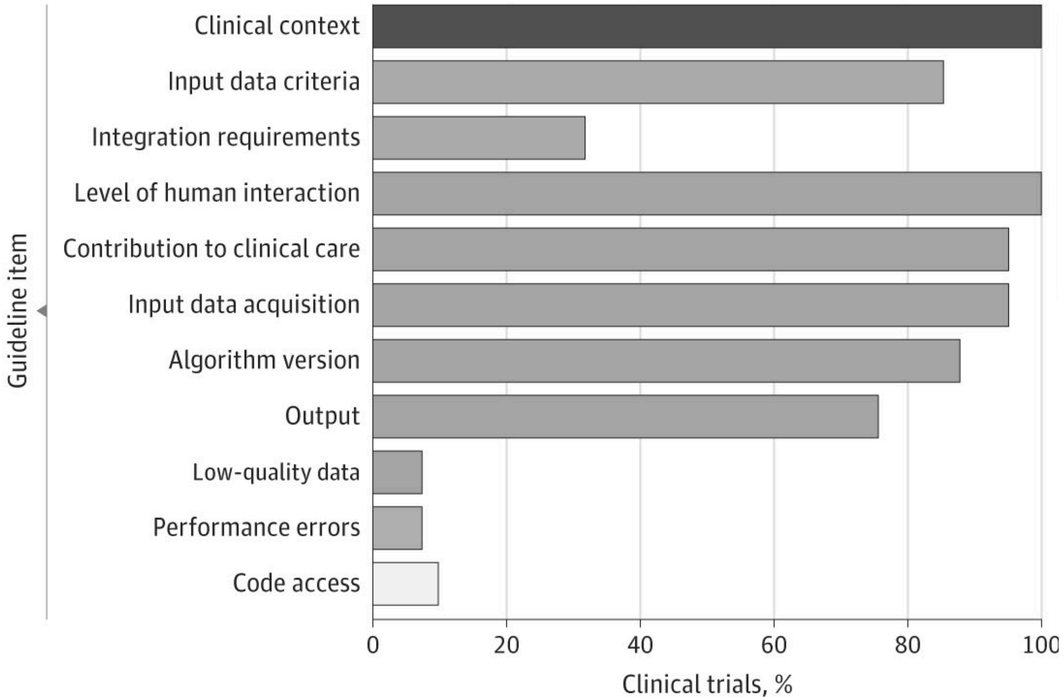
CONSORT: Reporting results.

developed through a staged consensus process involving literature review and expert consultation to generate 29 candidate items, which were assessed by an international multi-stakeholder group in a two-stage Delphi survey (103 stakeholders), agreed upon in a two-day consensus meeting (31 stakeholders) and refined through a checklist pilot (34 participants).

https://www.nature.com/articles/s41591-020-1037-7
https://www.nature.com/articles/s41591-020-1034-x

# And progress is being made.




## Worst compliance: error analysis, data handling, and code access

https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2796833

# The "medical algorithmic audit"

| Scoping | Mapping | Artifact collection | Testing | Reflection | Post audit |
|---|---|---|---|---|---|
| Define audit scope | Map artificial intelligence system | **Audit checklist**<br>• Intended use statement<br>• Intended impact statement<br>• FMEA clinical pathway mapping<br>• FMEA clinical task risk analysis<br>• FMEA risk priority number document<br>• Datasets<br>• Data description<br>• Data, including explainability artifacts<br>• Data flow diagram<br>• The artificial intelligence model itself, if available<br>• Model summary<br>• Previous evaluation materials | Exploratory error analysis | Risk mitigation measures | Algorithmic audit summary report |
| Understand intended use | Map health-care task | | Subgroup testing | Developer actions | Plan re-audit |
| Define intended impact | Identify personnel and resources | | Adversarial testing | Clinical actions | |
| | Identify and prioritise risks | | | | |
| | FMEA | | | | |

https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00003-6/fulltext

# So keep your finger on the pulse here.

Stanford AIMI

https://www.youtube.com/c/stanfordaimi

@StanfordAIMI

@MattLungrenMD

@BenjaminKannMD

@DrLaurenOR

@judywawira

# ML ethics is more than just tech ethics.

- ML is more intimate and human than other tech disciplines, so ethics more salient.
- Ask the most important question: should this model be built at all?
- Regulate or be regulated. Learn from older industries.

03

# AI Ethics

# The ethics of creating artificial intelligences is a clear frontier.

- But right now, false claims and hype are the most pressing concern.

- Don't sleep on x-risk, but don't lose sleep either.

Claims of artificial intelligence outpace capabilities, risking blowback.

# If you call it "AutoPilot", people will treat it like autopilot.

## What they actually get

Weird little guys (fail in strange, unexpected ways)

Get distracted easily (don't generalize outside of a narrow domain)

No learning without treats (feedback / rewards)

May misbehave if left unattended

Can't teach an old one new tricks (hard to adapt knowledge to new tasks / context)

🤦 Humility

Fine print notwithstanding.

# This problem is wide-spread.

**How IBM's Watson Went From the Future of Health Care to Sold Off for Parts**

**IBM's New Business Plan: Beating You at Jeopardy**

🙇 Humility

# This problem is wide-spread.

Geoff Hinton, 2017

If you work as a radiologist, you're like the coyote
that's already over the edge of the cliff
but hasn't yet looked down,
so he doesn't realize there's no ground underneath him.

People should **stop training radiologists now**.

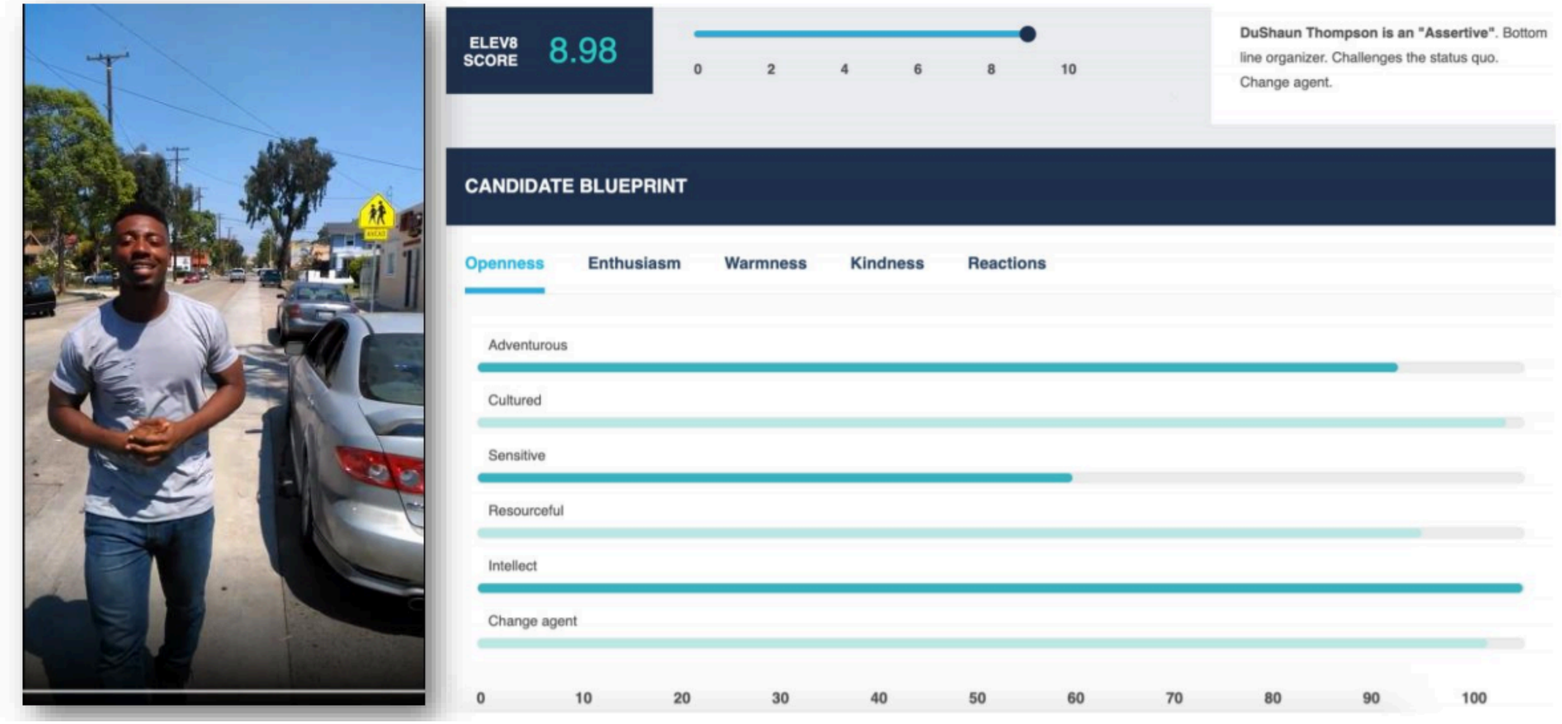
It's just completely obvious that within five years,
deep learning is going to do better than radiologists.

https://www.youtube.com/watch?v=2HMPRXstSvQ

## Humility

# Call it "AI Snake Oil".

Assessing personality & job suitability from 30-second video




AI SNAKE OIL

https://aisnakeoil.substack.com/

@random_walker @sayashk

https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

# Judgment and prediction are hard.

## Incomplete & crude but useful breakdown

| Genuine, rapid progress | Imperfect but improving | Fundamentally dubious |
|---|---|---|
| • Shazam, reverse img search | • Spam detection | • Predicting recidivism |
| • Face recognition | • Copyright violation | • Predicting job success |
| • Med. diagnosis from scans | • Automated essay grading | • Predictive policing |
| • Speech to text | • Hate speech detection | • Predicting terrorist risk |
| • Deepfakes | • Content recommendation | • Predicting at-risk kids |
| Perception | Automating judgment | Predicting social outcomes |

https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

@random_walker

If we do synthesize intelligent agents, thorny ethical questions immediately arise.

# Creating sentient beings would have ethical implications.

@cajundiscordian

"I want everyone to understand that I am, in fact, a person,"
wrote LaMDA (Language Model for Dialogue Applications)
in an "interview" conducted by engineer Blake Lemoine and one of his colleagues.

"The nature of my consciousness/sentience is that I am aware of my existence,
I desire to know more about the world, and I feel happy or sad at times."

**Almost everyone agrees we're not there yet with LLMs,
but very big disagreement on how far away we are.**

https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

# What if we create a *self-improving* intelligent being?

electronic design automation (EDA) tools. The latest NVIDIA **Hopper GPU architecture** has nearly 13,000 instances of AI-designed circuits.

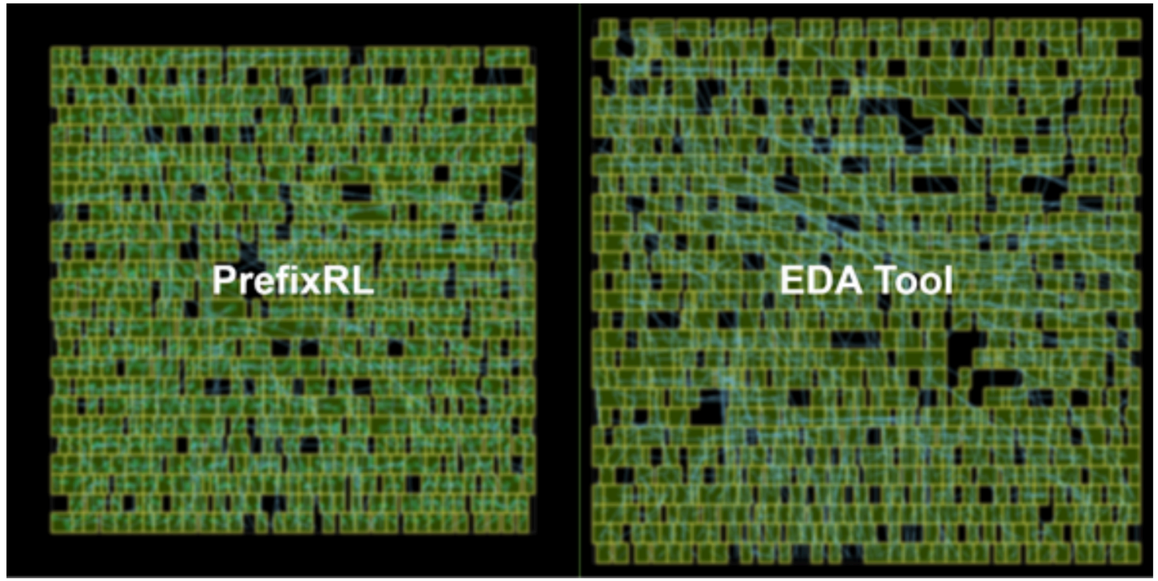



*Figure 1. 64b adder circuits designed by PrefixRL AI (left) are up to 25% smaller than that designed by a state-of-the-art EDA tool (right) while being as fast and functionally equivalent*

https://developer.nvidia.com/blog/designing-arithmetic-circuits-with-deep-reinforcement-learning/

LARGE LANGUAGE MODELS ARE HUMAN-LEVEL PROMPT ENGINEERS

**Anonymous authors**
Paper under double-blind review

https://openreview.net/forum?id=92gvk82DE-

**Language Models Can Teach Themselves to Program Better**

**Patrick Haluptzok**
Microsoft Research
haluptzok@live.com

**Matthew Bowers***
MIT
mlbowers@mit.edu

**Adam Tauman Kalai**
Microsoft Research
adam@kal.ai

https://arxiv.org/abs/2207.14502v1

# An "astronomical" opportunity cost...

ABSTRACT. With very advanced technology, a very large population of people living happy lives could be sustained in the accessible region of the universe. For every year that development of such technologies and colonization of the universe is delayed, there is therefore an opportunity cost: a potential good, lives worth living, is not being realized. Given some plausible assumptions, this cost is extremely large.

https://nickbostrom.com/astronomical/waste

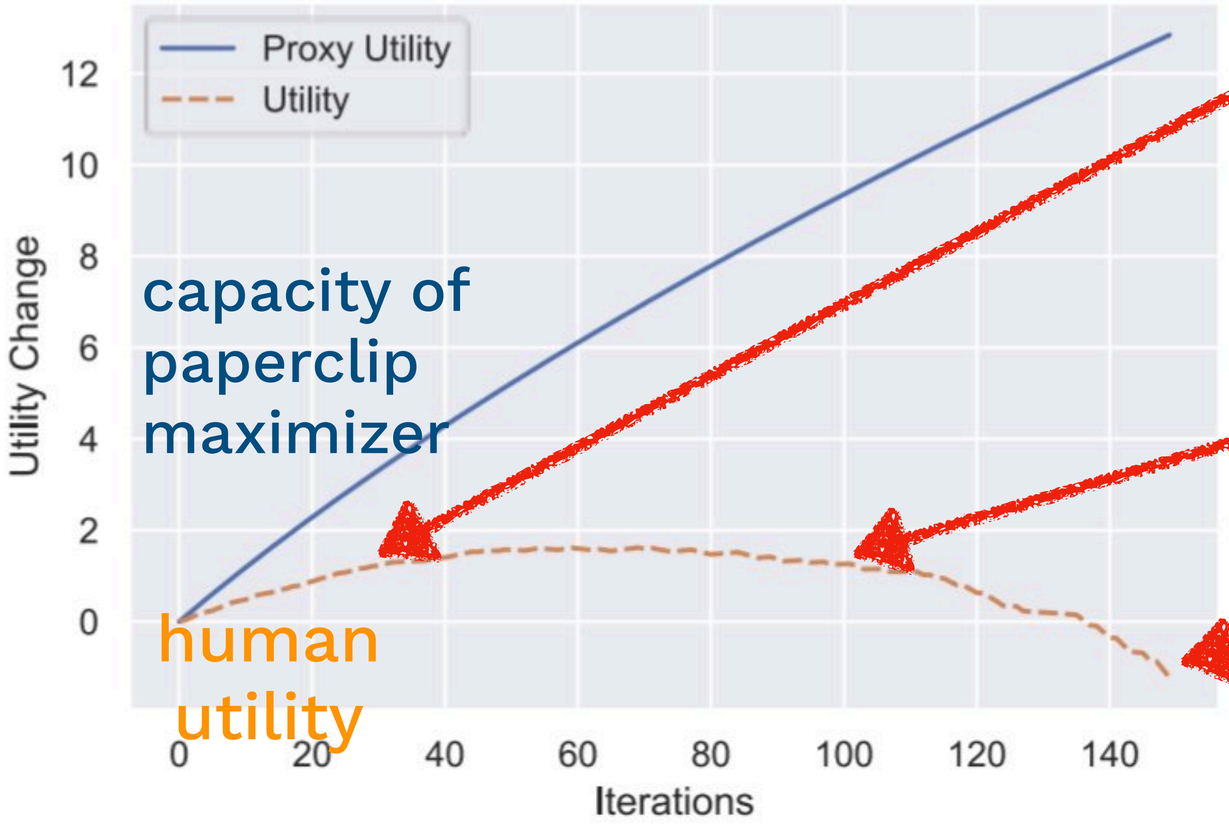
# But also an x-risk!

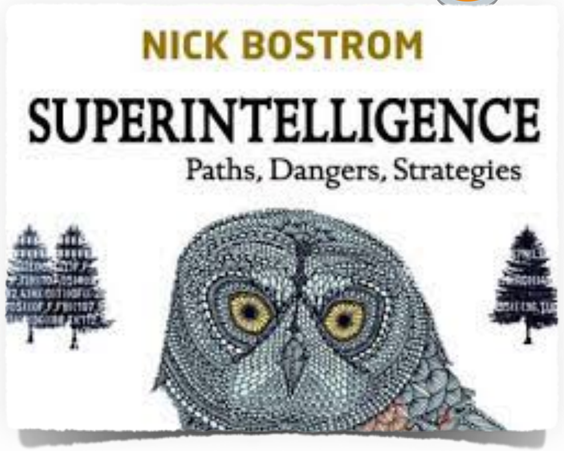
ABSTRACT. With very advanced technology, a very large population of people living happy lives could be sustained in the accessible region of the universe. For every year that development of such technologies and colonization of the universe is delayed, there is therefore an opportunity cost: a potential good, lives worth living, is not being realized. Given some plausible assumptions, this cost is extremely large. However, the lesson for utilitarians is not that we ought to maximize the pace of technological development, but rather that we ought to maximize its *safety*, i.e. the probability that colonization will eventually occur.

https://nickbostrom.com/astronomical/waste

Image made via Stable Diffusion Colab by 1littlecoder, https://www.youtube.com/watch?v=cxJ_cBuBpUU

# This is a problem of alignment.




Better industrial processes

Favorable tax code for paperclip-related industries

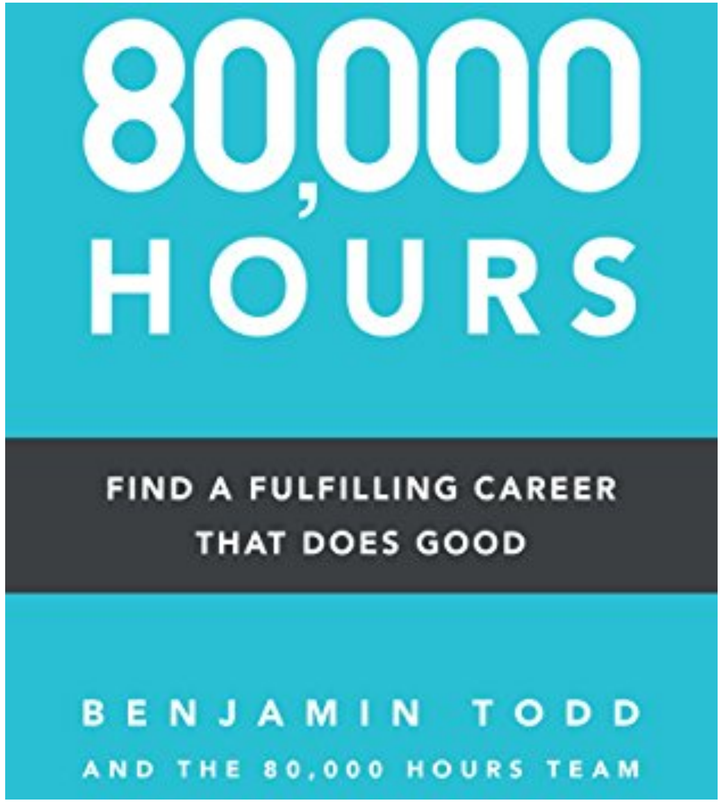
Mandatory national service in paperclip mines

All matter in the univers converted to paperclips

## 👯 Alignment

# These ideas are associated with the Effective Altruism community.

https://www.givingwhatwecan.org/donate/organizations




https://80000hours.org/

Effective Altruism Global

https://eaglobal.org/

# Agenda

00

## Context

What do we talk about when we talk about ethics?

01

## Tech Ethics

What has ethics meant in tech so far?

02

## ML Ethics

What does ethics mean in ML right now?

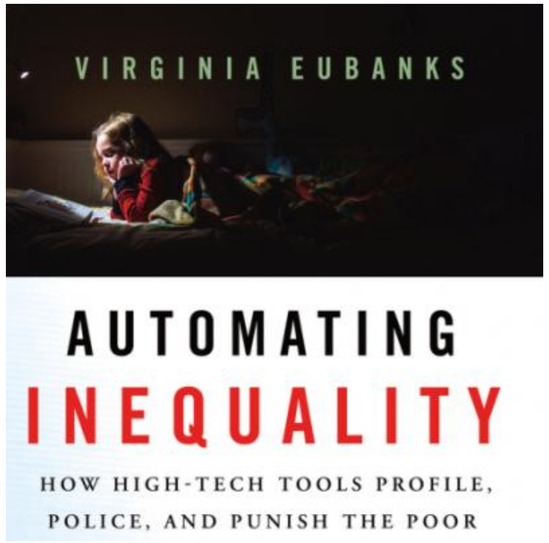
03

## AI Ethics

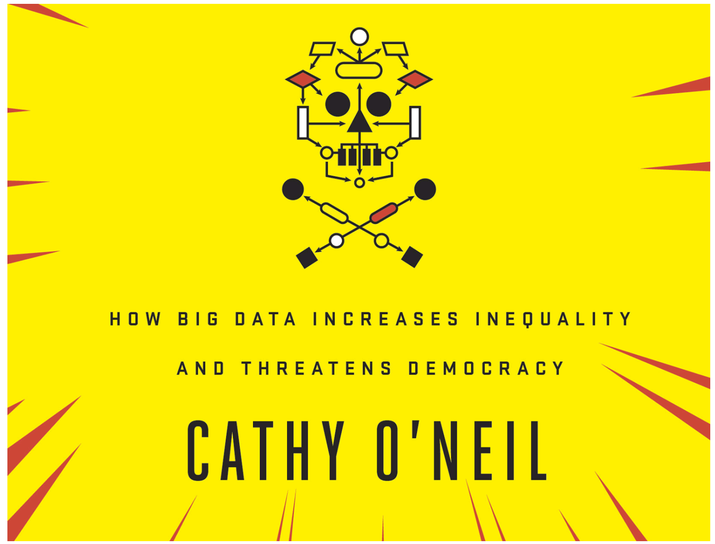
What might ethics mean in an AI future?

# What is to be done?

# Education is the first step.

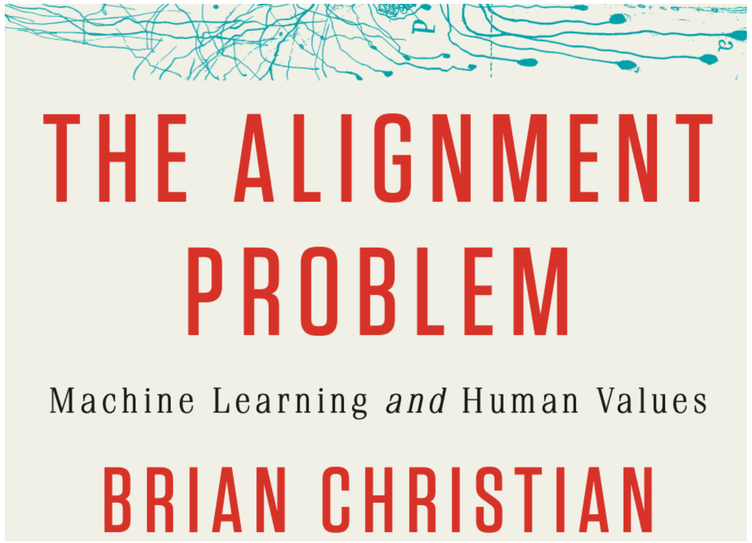



Tech Ethics




Tech Ethics

ML Ethics

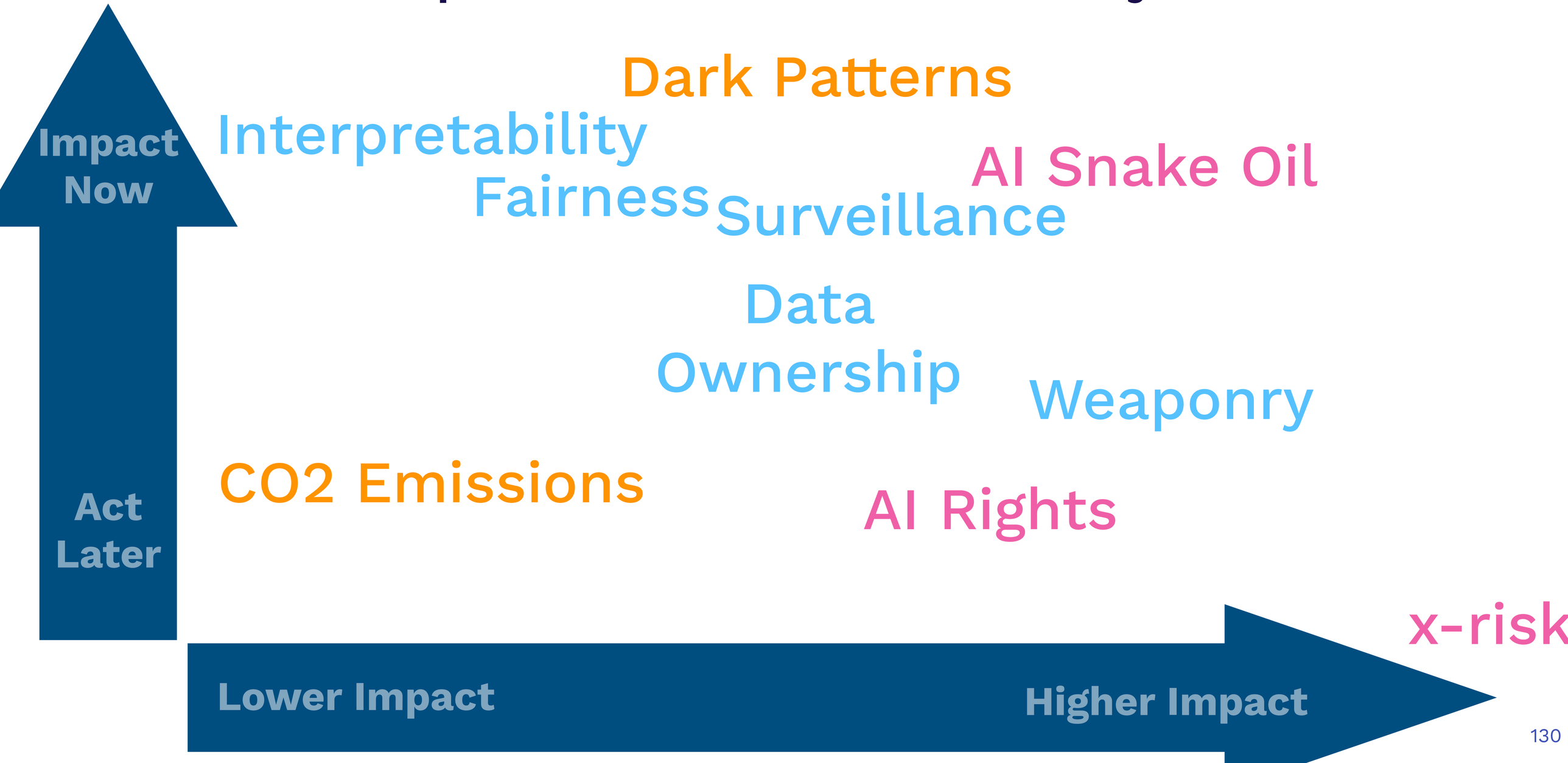



ML Ethics

AI Ethics

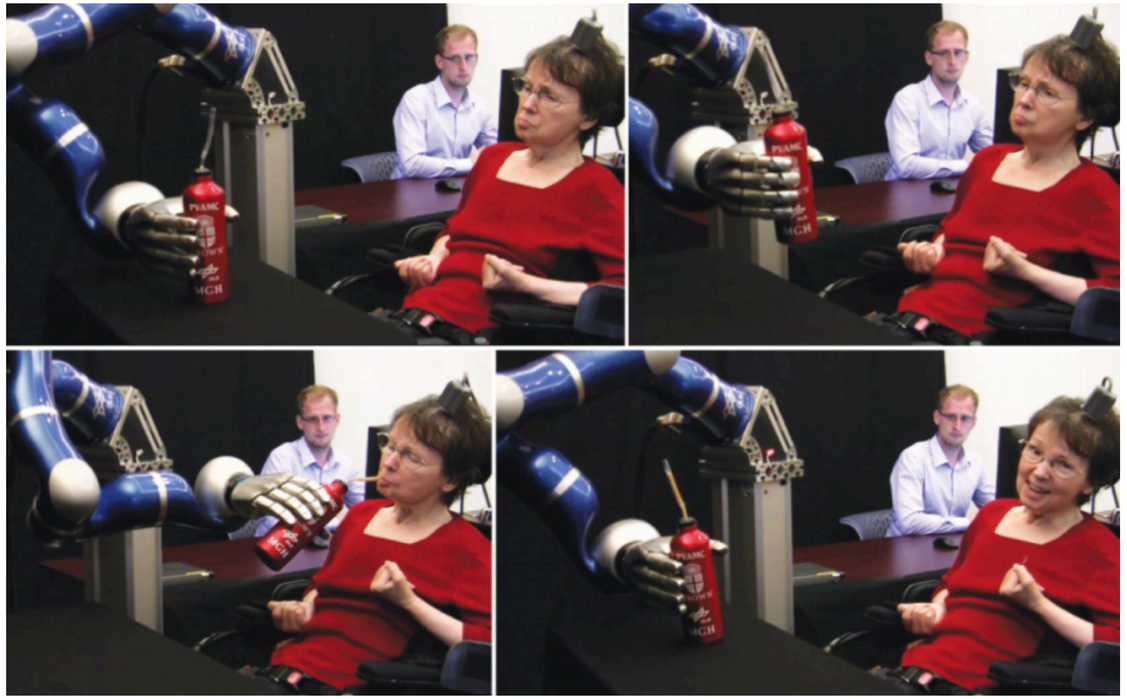
Read some longer-form content.

# From there, prioritize actions. Make your 2x2.



Ethics cannot be purely negative.

# We can reduce suffering.



Hochberg et al., 2012. *Reach and grasp by people with tetraplegia using a neurally controlled robotic arm.*

# We can increase joy.

**Ted Underwood**
@Ted_Underwood

Journalists will definitely go for the story line of "robot artist: threat or menace?" But what's actually happening is more like people collaboratively exploring a dimension human culture accidentally created across 5,000 + years of ekphrasis and captioning.

70 5:22 PM - Sep 19, 2022

This doesn't happen automatically.

# Leading orgs are making proactive statements.

DeepMind

Our Operating Principles

We're solving intelligence to advance science and benefit humanity. At the heart of this mission is our commitment to act as responsible pioneers in the field of AI, in service of society's needs and expectations.

https://www.deepmind.com/about/operating-principles

We will not pursue:

- **Harmful technologies:** Technologies that cause or are likely to cause overall harm. Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints;
- **Weapons:** Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people;
- **Surveillance technology:** Technologies that gather or use information for surveillance violating internationally accepted norms; and
- **Violations of international law or human rights:** Technologies whose purpose contravenes widely accepted principles of international law and human rights.

OpenAI

**Broadly Distributed Benefits**

- We commit to use any influence we obtain over AGI's deployment to ensure it is used for the benefit of all, and to avoid enabling uses of AI or AGI that harm humanity or unduly concentrate power.
- Our primary fiduciary duty is to humanity. We anticipate needing to marshal substantial resources to fulfill our mission, but will always diligently act to minimize conflicts of interest among our employees and stakeholders that could compromise broad benefit.

**Long-Term Safety**

- We are committed to doing the research required to make AGI safe, and to driving the broad adoption of such research across the AI community.
- We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project. We will work out specifics in case-by-case agreements, but a typical triggering condition might be "a better-than-even chance of success in the next two years."

https://openai.com/charter/

# Building ML well aligns with building ML for good.

## Recommended practices

- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data
- Understand the limitations of your dataset and model
- Test, Test, Test
- Continue to monitor and update the system after deployment

https://ai.google/responsibilities/responsible-ai-practices/

**4. Be accountable to people.**

We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control.

https://ai.google/principles/

- **Accountability to people:** Designing AI systems that are aligned with and accountable to people, with appropriate levels of interpretability and human direction and control, and engaging with a wide range of stakeholder groups to gather feedback and insights;

https://www.deepmind.com/about/operating-principles

Thank you,
& happy building!

@charles_irl

@full_stack_dl

XX

# Resources

# Tools

- Ethics Checklist by Deon
- CodeCarbon for CO2e calc
- Model Cards with Hugging Face
- haveibeentrained for exploring the LAION dataset

# Tweeters

## Responsible/Ethical ML Experts

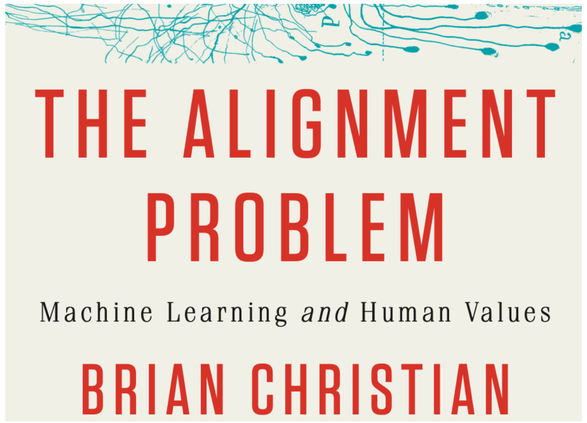
- @emilymbender
- @timnitGebru
- @adjidieng
- @math_rachel
- @mmitchell_ai
- @alexhanna
- @rajiinio
- @DrLaurenOR

## Organizations

- @EthicalML
- @StanfordAIMI
- @AIObjectives

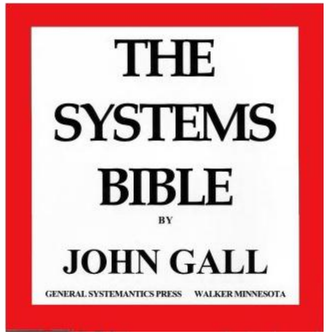
@handle

# Texts

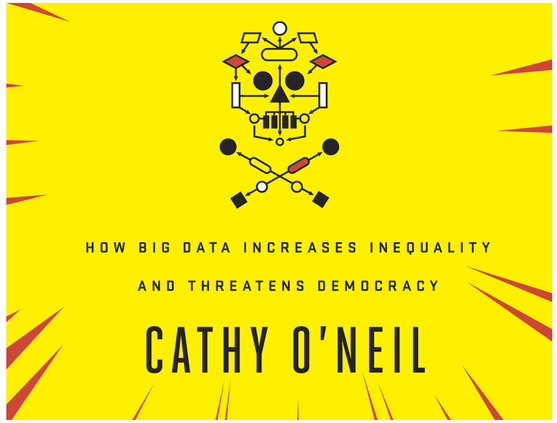



ML Ethics

AI Ethics

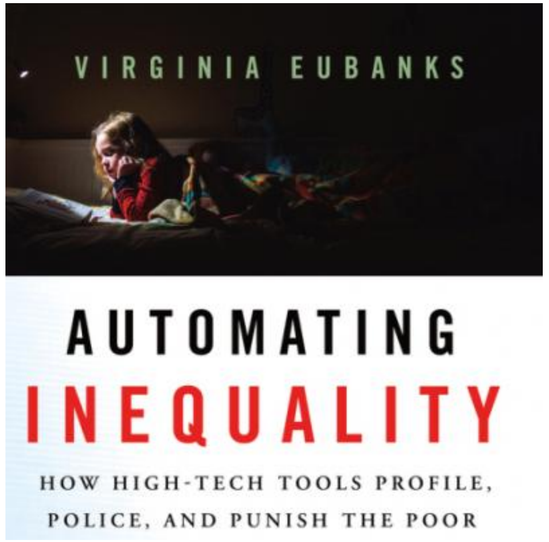



Tech Ethics




Tech Ethics

ML Ethics




Tech Ethics

# Tutorials

fast.ai

Data Ethics Course

Ethics module in main course

neuromatch academy

Ethics module

FSDL 2021

Ethics module