Week 7: Ethics i

Charlotte Hadley

Ethics in algorit
it mean?

Ethics in algorithms mean?

"Ethics in algorithms" is a part of data ethics that we only need to consider when we're discussing writing algorithms.

Last week we discussed data anonymisation (and re-identification) a lot of our conversation was about the ethics of data collection.

> Is it ethical for data about us to be collected without our knowledge?

Data ethics includes,

- How data is collected
- How data is processed
- How data is consumed and services
- How data is released

Companies are hiring droves.

Data ethics… you mean machine le

Machine learning, big data and AI are all tremendously exciting and de always require data.[1]

But data ethics is important every time we're dealing with sensitive and/or private data.

Even if we've got survey data about attitudes to green spaces during the pandemic.

In this course when we talk about data ethics we're also talking about the ethics of algorithms.

Ethics & Moral

Ethics & Moral Philo

Moral philosophy and the history of ethics is fascinating.

We're just going to it.

R i g h t   t o   p r i v

Right to privacy

The Universal Declaration of Human Rights[2] provides universal right to privacy.

Article 12

No one shall be subjecte

interference with his home or pri

correspondence, nor to at

and reputation. Everyone

protection of the law aga

or attacks.

We're going to extend this to include right to privacy in data collection and sharing of data.

But let's look at algorithms a little bit more

Self-driving

Self-driving cars (I

There's a continuum of "self-driving car" technology. For discussion purposes, let's define the tech as follows:

> A car on public roads under the control of an automated driving system that controls acceleration, breaking and driving direction. The driver does not have active control of the car, but can disengage the automated driving system and resume control.

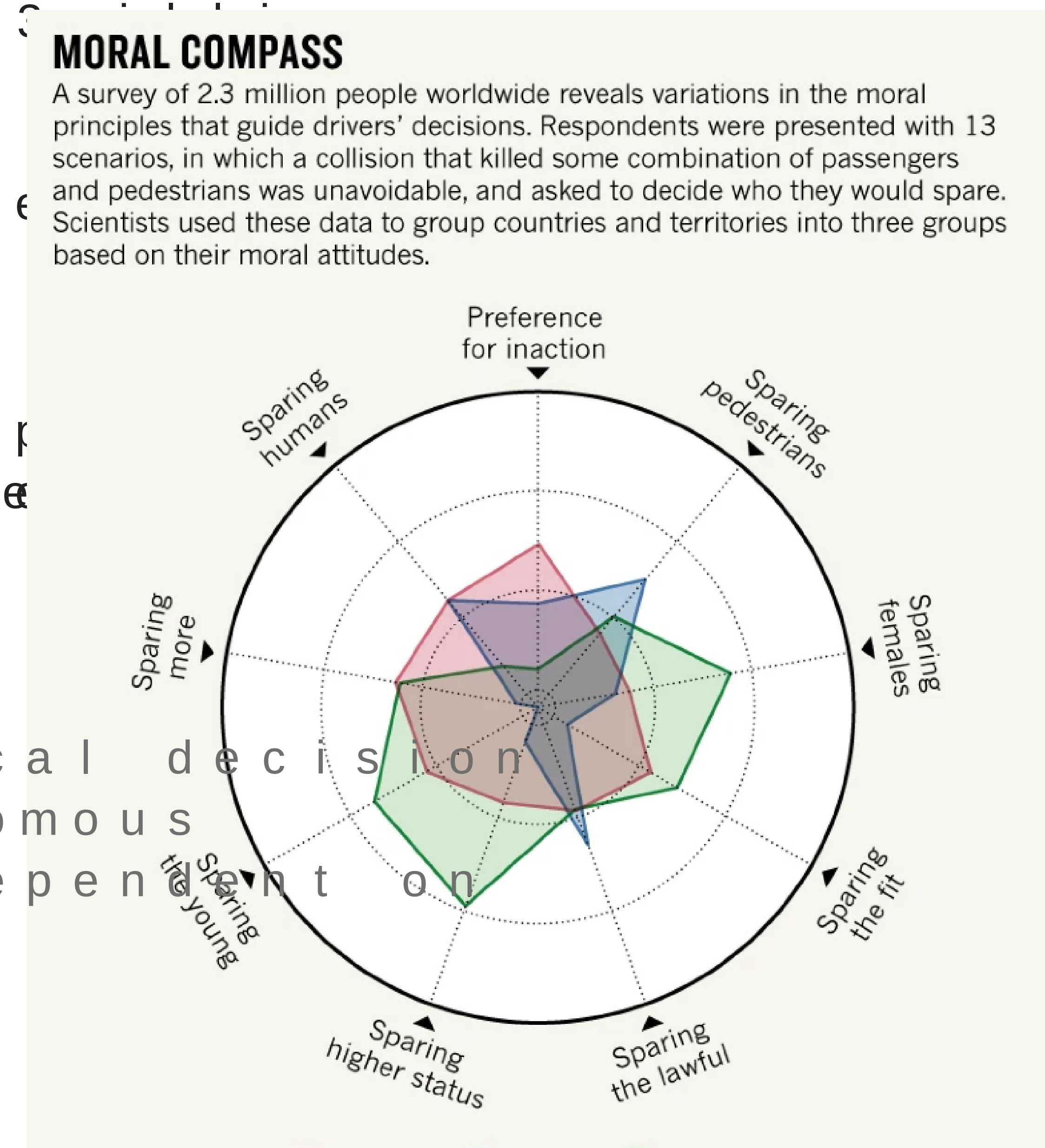
What ethical questions does the addition of self-driving cars to existing road networks raise?

- Who is responsible for a self-driving car collision?
- How does the automated respond to trolley p
  - Does the car priori the driver or a ped
  - Does the car priori other vehicles or p

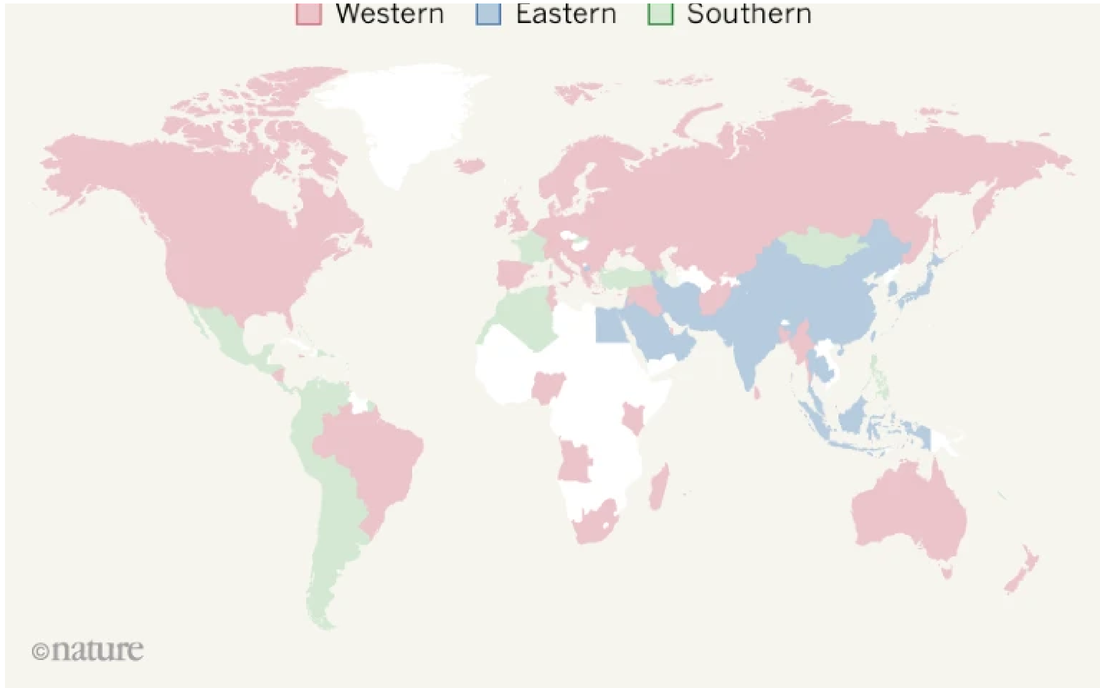
Self-driving cars (I

Awad et al.[3] surveyed 2.3
people in 2018 to
to moral dilemmas
autonomous vehicle

Attitudes vary by
region of partici
demographics of pe
victims



> Should the ethical decision making of autonomous vehicles vary dependent on location?!



This data visualisation is derived from Awad et al[3] but appears in Maxmen 2018[4]

Self-driving cars (I

Reliable and unbiased estimates of the vehicle safety are difficult to find. This is not a totally fair comparison though, interview...

As more autonomous vehicles are used by untrained drivers in real-world circumstances, collect more information about their safety.

Self-driving car[s have] an accident rate [of] crashes per mil[lion miles] driven – which i[s] double that of r[egular] vehicles.

It's derived from a ... and a total of ...

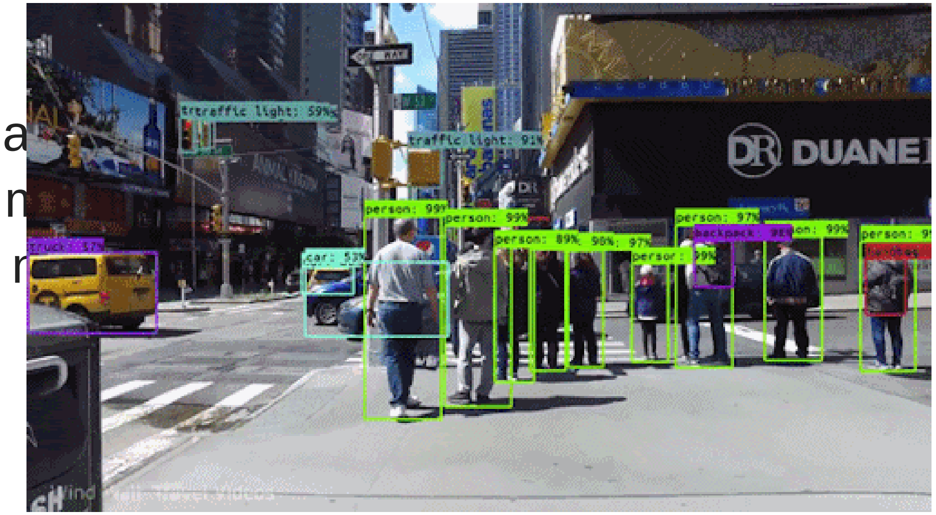
But it's important we throw and understand autonomous vehicles make their decisions.

Self-driving cars (l

Self-driving vehicles use many tools for sensing and environment.

There's considerable deba[te and dispute] about two com[peting] technologies: LiDAR and Vision.

- Computer vision uses and categorizes what measuring other



- LiDAR continuously pu[lses] measures reflection times and 3D visual map. This c[an be] clustered and classified to identify around the car.

Self-driving cars (I

As we question the eth
autonomous svechriuccliesal itt
undershtoawnvdehicles make
particularly during ac

Does the vehi,cAecmoakte
andr aTnspaATe)ntde(cFisions?

Thank,fublolyth LiDAR and
provide surprriasnipngalreynk

This table comes from
rep[6]obetween an autonon
a pedestrian.

Time to Impact (seconds)	Speed (mph)	Classification and Path Prediction[a]	Vehicle and System Actions[b]
-9.9	35.1	--	Vehicle begins to accelerate from 35 mph in response to increased speed limit.
-5.8	44.1	--	Vehicle reaches 44 mph.
-5.6	44.3	Classification: *Vehicle*—by radar Path prediction: *None*; not on path of SUV	Radar makes first detection of pedestrian (classified as vehicle) and estimates speed.
-5.2	44.6	Classification: *Other*—by lidar Path prediction: *Static*; not on path of SUV	Lidar detects unknown object. Object is considered new, tracking history is unavailable, and velocity cannot be determined. ADS predicts object's path as static.
-4.2	44.8	Classification: *Vehicle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *vehicle*; this is a changed classification of object and without a tracking history. ADS predicts object's path as static.
-3.9[c]	44.8	Classification: *Vehicle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains classification *vehicle*. Based on tracking history and assigned goal, ADS predicts object's path as traveling in left through lane.
-3.8 to -2.7	44.7	Classification: alternates between *vehicle* and *other*—by lidar Path prediction: alternates between *static* and left through lane; neither considered on path of SUV	Object's classification alternates several times between *vehicle* and *other*. At each change, tracking history is unavailable; ADS predicts object's path as static. When detected object's classification remains same, ADS predicts path as traveling in left through lane.
-2.6	44.6	Classification: *Bicycle*—by lidar Path prediction: *Static*; not on path of SUV	Lidar classifies detected object as *bicycle*; this is a changed classification of object and object is without a tracking history. ADS predicts bicycle's path as static.
-2.5	44.6	Classification: *Bicycle*—by lidar Path prediction: Left through lane (next to SUV); not on path of SUV	Lidar retains *bicycle* classification; based on tracking history and assigned goal, ADS predicts bicycle's path as traveling in left through lane.

We can see the autonomous vehicle
repea,teadnldy inconsisftdently misidenti
a pedestrian.

Fairness, Account
Transparency

Fairness, Accountabi
Transparency

Fairness

Fairness: 5 (or 6) S

Frustratingly, the source for this! But this list has been around since at least[8] 2014.

Broadly in the literature (and in essays on the subject) there are 5 common sources of bias in training datasets.

These biases result in unfair treatment of individuals/groups by an algorithm.

Examples of unfair treatment include

- Refusal of service
- More expensive services
- Reduced range of services

When these services include healthcare the ramifications can be life threatening.

- Proxies
- Limited features
- Skewed sample
- Tainted examples
- Sample size disparit

The assessment limits sources.

Is it worthwhile mention these bias, acknowledging left (or added) to a

This is masking.[8]

Fairness: Proxies

> Proxies are the easiest to identify and d

When training our algorithms we want to rem
gender, age or religious belief - because we d
disadvantage or advantage a specific group

However, there are many other variables that are proxies for these t
Can you think of any?

- Systemic racism means that income, neighbo
  proxies for race.

> "Redlining is the practice of financial institutions
> neighborhoods, generally because its" resi

- Facebook friendships can be a[11] strong proxy

Fairness: Limited Fe

> This is a harder bias to address extremely (and so pr
> down examples

Limited features is a consequence of having
specific combinations of sensitive attributes.

Individuals in these groups will be less (or more
than other groups.

This article gives a theoretical example: h
unintended-bias-in-training-data-be5b7f3347

> Limited features is a pre-requisite for s

Fairness: Skewed sam

> It's kind of unfair this is collected togeth
> examples, sample size disparities

Skewed samples is a bias in your algorithm
results in a becomes the skewed recovering b
feedback[12]. Please do read the paper[12] Ensign from 2017 as
really well.

> Given historical crime incident data for
> patrol officers to areas to detect crime
>
> ...
>
> Since such discovered incidents only occu
> sent to by the predictive policing algori
> bias to be compounded, causing a runaway

Fairness: Skewed sam

Another really great example with an easy
issues with Google's predictive power, Lazer et al 2014[13]

In 2009 Google published details of a model (GFT)[14] [illegible]
that used search results to predict flu outbreaks.

The theory being that ill people search for
detected.

However, the whole thing is somewhat questiona

GFT has never documented the 45 search te
been released appear misleading[15]

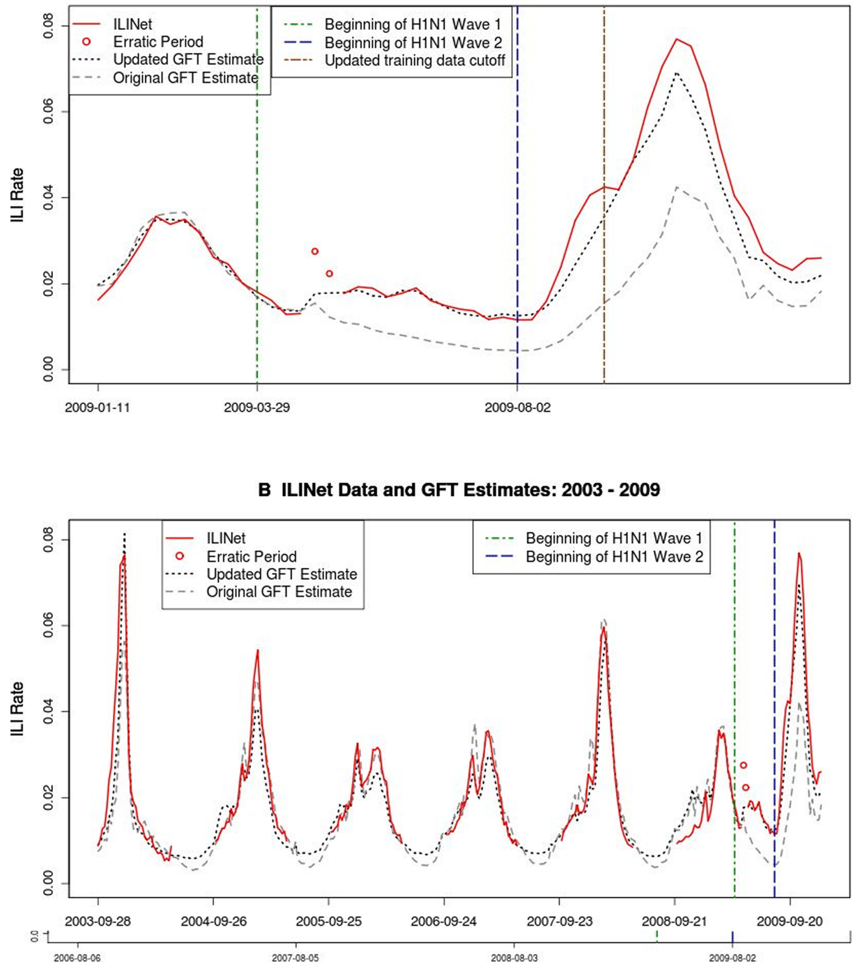
Source: Lazer et al 2014[13]

Fairness: Skewed sam (II)

Coincidentally first big fail
service was the 2009[15]

The GFT didn't detect t
wave of H1N1 during t .

Google quickly update hm
later in 2009 and the hm d
a better job of predi k.

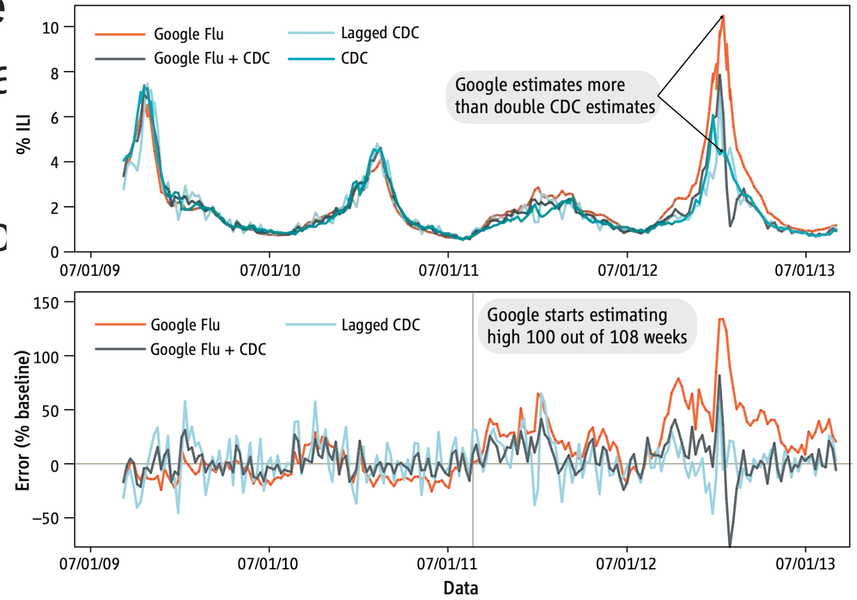


Fairness: Skewed sam (III)

However, from 2011 onwards the GFT
consistently overestimated flu [13]
and became increasing
Lazer[13] points to two reasons

- Big data hubris

Was the algorithm a flu or winter detector?!



Source: Lazer et al [13]

- Algorithm dynamic

Google has many competing... The algorithm has a pre...
that shows potentially interesting (or use...

Fairness: Tested examples

We need to talk ab
learning.

Do you know the db
supervised and uns

In supervised learn
categories and ass
these categories.

The algorithm will
as belonging to one

That's all well and g
something like the tita
survival dataset

… but often our ca
absolute, eg

- Google uses flu! has t
- Rating a driver's likelihoo
  accident
- Rating a "good jo

Let's look into the
example more

Fairness: examples in …

Algorithms are used sometimes aggressively in recruitment. Fraij and László published a literature review on the topic in 2021[16]

These algorithms are often marketed as software to automatically categorising "good" CVs.

The algorithms are celebrated for being efficient and allowing recruiters to focus on more interesting and human required tasks.

- Who gets hired. And … "shocking levels of … hiring"[17]
- Work evaluation scores … people's … evident discrimination in the …[18]

> This couldn't possibly create issues

Fairness: Sample Size Dis

> Unfortunately even perfectly balanced [untainte
> untainted categories can have bias. Sampl
> this.

Sample size disparity Tehxei smtoss tw hweindely cited
completely balanced daetxaasmeptlse hoafv et his i1s t h
disparity in the size hotft pssu:b/g/reonu.pwsi.kipe.dia.

This disparity results in unfair behaviour
the algorithm when applied to real-world
datasets.

Fairness: Sample Size Dis

In 2011 Google implemented a real name policy for Google+.

There are significant privacy issues with this policy.

But users also found their names rejected as “not real”.

Folks with names that don't fit the first name last name pattern - e found their names rej

This is widely blamed datasets having sampl there were fewer inst forms.

I recommend reading t Falsehoods Programmer Names – With Examples[19].

… some more cas

“Man is to computer programmer a… (I)

In 2016 Bolukbasi et al.[20] demonstrated how the extremely common word2vec algorithm shows gender bias.

For example, if you … distance will displa… words and their dist… which should look li…

word2vec is a patented[21] algorithm that – when trained on a given tex… - will find similar/synonymous words … given input.

Google provides a tutorial on how the system works.

"Man is to computer programmer a
(II)

"Man is to computer programmer a… (III)

What sources of bias are included in this example?

- Proxies
- Limited features
- Skewed sample
- Tainted examples
- Sample size disparit…

Racial bias in photo

Smartphone manufacturers have recently started advertising how t... accurately photograph the ... ll skin tones.[22]

This has known been an issue in photography, for decades in many different situations.

In pre-digital photograph... k issued "Shirley cards" fo... ancing images. These cards exclu... contained Caucasian models until the 1990s.

Figure 1
Polaroid Shirley card (Printed with permission of Polaroid)

Source: Lorna Roth[23]

Racial bias in photo

This issue has also been pervasive in digital photography.

TABLE 1 Fitzpatrick Classification of Skin Types I through VI

Type I	Type II	Type III	Type IV	Type V	Type VI
White skin. Always burns, never tans.	Fair skin. Always burns, tans with difficulty.	Average skin color. Sometimes mild burn, tan about average.	Light-brown skin. Rarely burns. Tans easily.	Brown skin. Never burns. Tans very easily.	Black skin. Heavily pigmented. Never burns, tans very easily.

The continued use of the Fitzpatrick skin tone scale is largely to blame.

Source: Ward et al[25]

Dr Ellis Monk's research culminated in the Monk Skin Tone Scale[24] with evidence improvements in processing of images compared to the Fitzpatrick scale.

Racial bias in photo

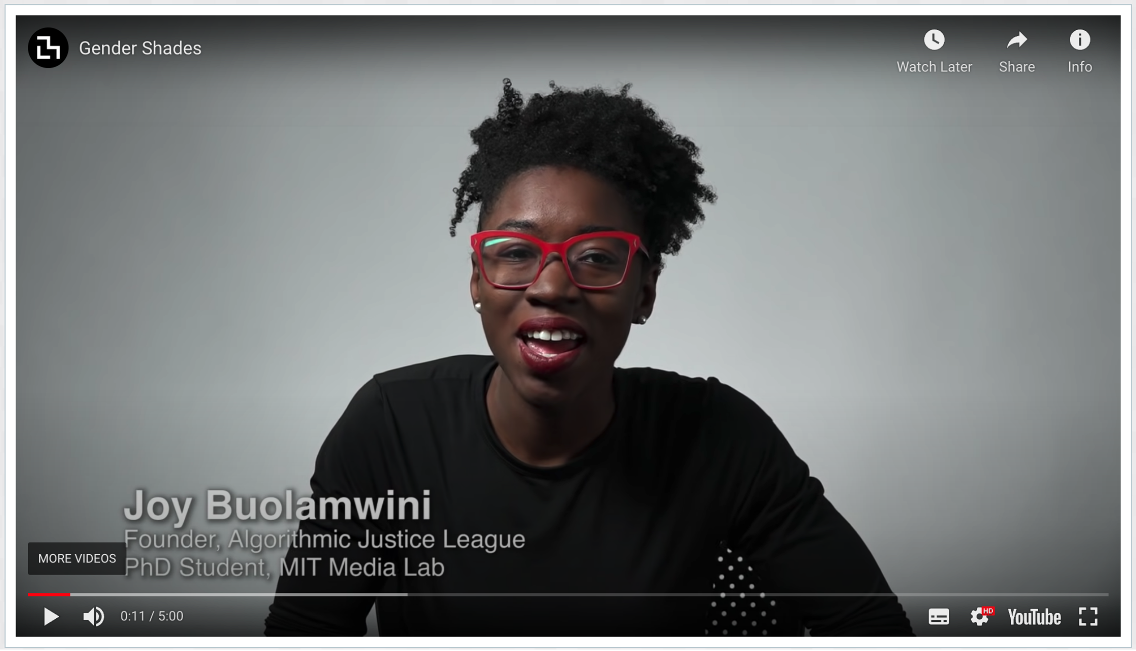
I'd like to recommend some additional resou

- Vox Media's video Color film was built for white people
  skin.[26]
- The Racial Bias Built Into Photography by[27]
- Lorna Roth's research paper, Looking at Shirley, the Ultimate Norm:
  Image Technologies, and Cognitive Equity[28]
- Google's Skin Tone Research with Dr. Ellis Monk.

Racial bias in face

This is a fundame
all face analysis
software.

Joy Buolamwini does a mu
better job of exp
minutes than I ca



Source: gendershades.org 28

Racial bias in face photography

What sources of bias are included in this example?

- Proxies
- Limited features
- Skewed sample
- Tainted examples
- Sample size disparit

Accountability & Transparency

Accountability & Transparency

These two terms are used interchangeably as … discuss how these two concepts interrelate to o…

These are really useful in the EU "governance framework … accountability[29] and transparency"

> The primary role of transparency is … not known what an organisation is doing, … regulated.
>
> …
>
> An important difference between transparency and a… is primarily a legal and ethical obligati… for its activities, accept responsibility … transparent manner

… but the rest of the document is very heav…

Accountability

For a more thorough exploration of acc[30]ounta

Transparency and gami

It's worthwhile mentioning that transparency algorithm is applied:

> Also, in some cases, transparency may lea
> system.” For example, even the minimal op
> feature Twitter surfaces topics has allowed i
> certain topics by bots and coordinated gr
> contexts may really fall off of transparency
>
> Source: Caplan[30] et al 2018

- How else could we game algorithms?

Other sources of

Other sources of unf

Earlier we tried to ca Let's look at some of th
bias:

- Proxies
- Limited features
- Skewed sample
- Tinted examples
- Sample size disparities

These are almost always concerned with the training data behind algorithms.

There's a whole universe of ways we can bias an a how it's applied

Fairness and Abstraction Systems

Selbst et al published an excellent paper in 2019[31] on "abstraction traps".

These traps are designed to help us account for the interactions between the technical systems behind our algorithms and the social world in which they're applied.

I think they're really ll further context for o fairness.

The next several slid traps.

Selbst Abstraction Trap:

Failure to model the entire system over w
be enforced

Example:

In the US criminal justice pipeline algori
predict the "risk" of a defendant and dete

<br

However - usually this is the risk the defe
Reoffending is only occassionally considered

These risk assessment tools are presented c
not account for[31]) defendants consider (creation betwe

The intended appliction of the algorithm a
into account their use by judges and ther
measured

Selbst Abstraction trap

> Failure to understand how repurposing alg... context may be misleading, inaccurate, or ... different context

Here's a really nice quote

> Within computer science, it is considered ... be used for different contexts. "But what social context is," says Selbst. "A system design ... applied in Kentucky ... fairness. Or you can't have a system that ... then applied to employment. How we think ... totally different."
>
> Source: Karen Hao[32]

Selbst Abstraction: ap

Failure to account for the full meaning o be procedural, contextual, and contestabl mathematical formalisms

I did promise we would In algorithmic recrui philosophy is important for don't particularly mi moment. positives (candidates fit) because the interface t them out.

Fairness and discrimination are complex concepts that philosophers, sociologists, and lawyers have long debated. They are at times procedural, contextual, and politically contestable, and each of those properties is a core part of the concepts themselves[31]

However the justicial r mentioned earlier mig opposite treatment. F result in parole bein

Selbst Abstraction: Trap

> Failure to understand how the insertion o
> changes the behaviors and embedded values

Selbst[31] depend entirely on the justicial ris
increased use focuses the intent of the jus

Selbst Abstraction: Trap

- Failure to recognize the possibility that technology
- Modeling requires fixing definitions in code. When fairness is politically contested, movable issue, a m how it moves

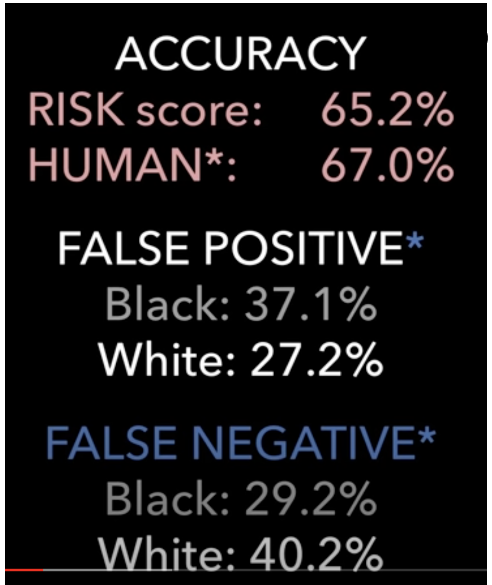
COMPAS: Algorithmi assessments

COMPAS: Algorithmic risk assessments

The algorithmic risk assessment algorithm we've discussed is called COMPAS.

It's been subject to scrutiny

ProPublica's study[33] in 2016 was one of the first investigations. A technical breakdown of this is provided by the same authors[34]



The algorithm performs as well as random Mechanical Turkers[35]

I'd like to show you Hany Farid's TED talk on this topic[36]

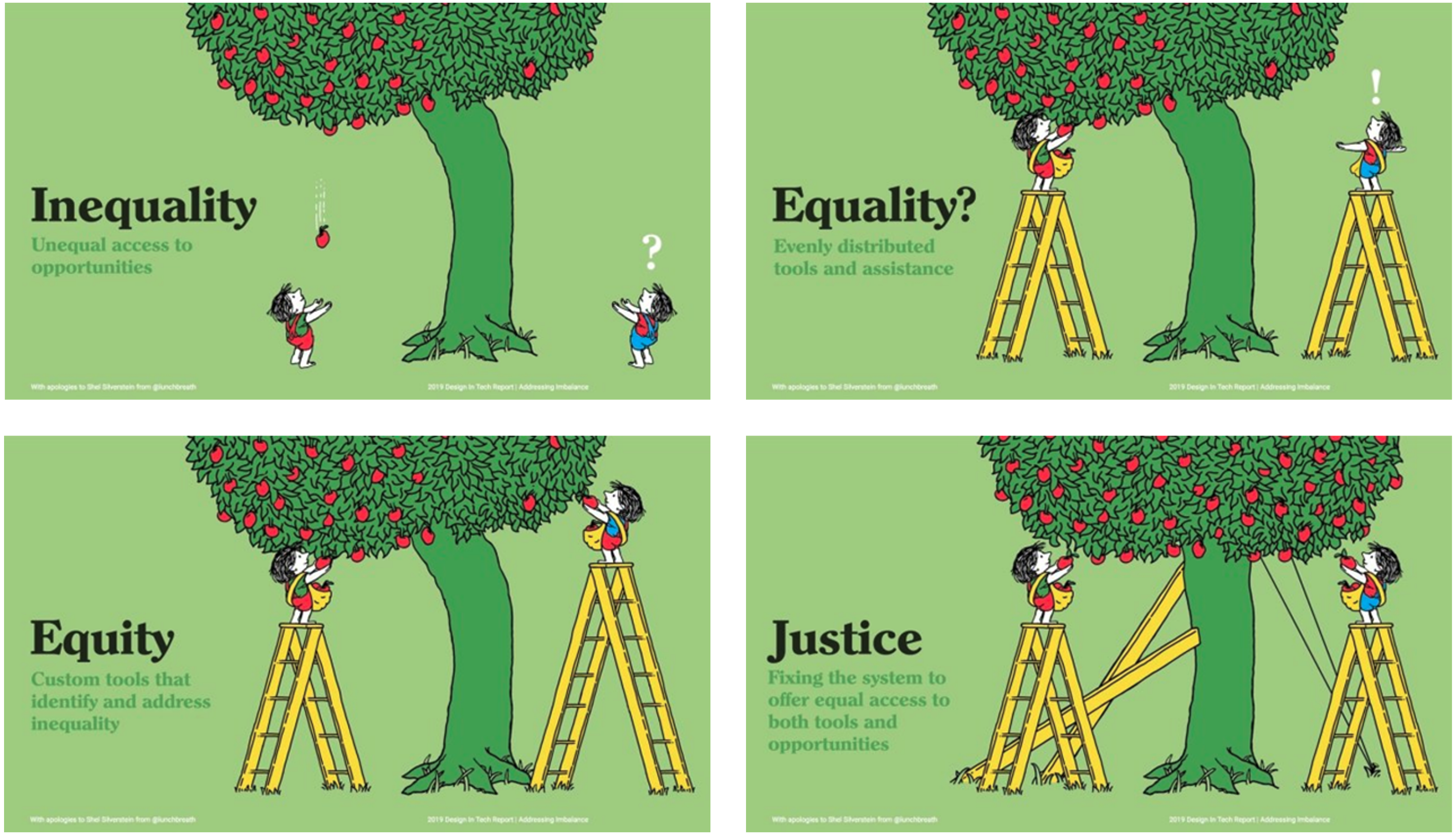
Inequality vs Jus

Inequality vs Justic

I thought it was important to not skip over

These cartoons are presented at John Maeda's Design in Tech Report[37] and 20

designed by Tony Ruth



We have to stop

We have to stop some

We've covered more than enough for your ass
complexities of ethics in algorithms.

We didn't really discuss data privacy very
Fair Information Practices.

References

1. Sucholutsky, I. & Schonlau, M. "Less Than One"-Shot Learning: Learning N Classes From M < N Samples. Proceedings of the AAAI Conference on Artificial Intelligence 35, 9739–9746 (2021).
2. United Nations General Assembly. Universal Declaration of Human Rights. (1948).
3. Awad, E. et al. The Moral Machine experiment. Nature 563, 59–64 (2018).
4. Maxmen, A. Self-driving car dilemmas reveal that moral choices are not universal. Nature 562, 469–470 (2018).
5. Schoettle, B. & Sivak, M. A Preliminary Analysis of Real-World Crashes Involving Self-Driving Vehicles. (2015).
6. National Transport Safety Board. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018. (2018).
7. Hardt, M. Occupy Algorithms: Will Algorithms Serve the 99%? (2013).
8. Barocas, S. & Selbst, A. D. Big Data's Disparate Impact. (2016) doi 10.2139/ssrn.2477899.
9. Landeau, A. Explaining Bias in Your Data. (2020).
10. Goldfain, C. Sources of unintended bias in training data. Medium (2020).
11. Jernigan, C. & Mistree, B. F. T. Gaydar: Facebook friendships expose sexual orientation. First Monday (2009) doi 10.5210/fm.v14i10.2611.
12. Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. Runaway Feedback Loops in Predictive Policing. (2017) doi 10.48550/arXiv.1706.09847.
13. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The Parable of Google Flu: Traps in Big Data Analysis. Science 343, 1203–1205 (2014).
14. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014 (2009).
15. Cook, S., Conrad, C., Fowlkes, A. L. & Mohebbi, M. H. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLOS ONE 6, e23610 (2011).
16. Frai, J. J. & László, V. A Literature Review: Artificial Intelligence Impact on the Recruitment Process. International Journal of Engineering and Management Sciences 6, 108–119 (2021).
17. Siddique, H. Minority ethnic Britons face 'shocking' job discrimination. The Guardian (2019).
18. Stauffer, J. M. & Buckley, M. R. The Existence and Nature of Racial Bias in Supervisory Ratings. Journal of Applied Psychology 90, 586–591 (2005).
19. rogers, tony. Falsehoods Programmers Believe About Names - With Examples. Shine Solutions Group (2018).

20. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. (2016) doi:10.48550/arXiv.1607.06520.
21. Mikolov, T., Chen, K., Corrado, G. S. & Dean, J. A. Computing numeric representations of words in a high-dimensional space. (2015).
22. Google. Real Tone on Google Pixel. Google Store (2022).
23. Roth, L. Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity. Canadian Journal of Communication 34, 111–136 (2009).
24. Monk, E. P., Jr. The Unceasing Significance of Colorism: Skin Tone Stratification in the United States. Daedalus 150, 76–90 (2021).
25. Ward, W. H., Lambreton, F., Goel, N., Yu, J. Q. & Farma, J. M. TABLE 1, Fitzpatrick Classification of Skin Types I through VI. (2017).
26. Vox. Color film was built for white people. Here's what it did to dark skin. (2015).
27. Lewis, S. The Racial Bias Built Into Photography. The New York Times (2019).
28. MIT Media Lab. Gender Shades. (2018).
29. European Parliament. Directorate General for Parliamentary Research Services. A governance framework for algorithmic accountability and transparency. (Publications Office, 2019).
30. Caplan, R., Donovan, J., Hanson, L. & Matthews, J. Algorithmic Accountability: A primer. (2018).
31. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. & Vertesi, J. Fairness and Abstraction in Sociotechnical Systems. in Proceedings of the Conference on Fairness, Accountability, and Transparency 59–68 (Association for Computing Machinery, 2019). doi:10.1145/3287560.3287598.
32. Hao, K. This is how AI bias really happens and why it's so hard to fix. MIT Technology Review (2019).
33. Mattu, L. K., Jeff Larson. Machine Bias. ProPublica (2016).
34. Mattu, L. K., Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica (2016).
35. Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. Science Advances 4, eaao5580 (2018).
36. TEDx Talks. The danger of predictive algorithms in criminal justice | Hany Farid | TEDxAmoskeagMillyard. (2018).
37. Maeda, J. Presentation: Design in Tech Report 2019. Design in Tech (2019).