Responsible Data Science

Legal Frameworks

May 3, 2022

Prof. Julia Stoyanovich

Center for Data Science &
Computer Science and Engineering
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

This week's reading

The imperative of interpretable machines

As artificial intelligence becomes prevalent in society, a framework is needed to connect interpretability and trust in algorithm-assisted decisions, for a range of stakeholders.

Julia Stoyanovich, Jay J. Van Bavel and Tessa V. West

We are in the midst of a global trend to regulate the use of algorithms, artificial intelligence (AI) and automated decision systems (ADS). As reported by the *One Hundred Year Study on Artificial Intelligence*: "AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy." Major cities, states and national governments are establishing task forces, passing laws and issuing guidelines about responsible development and use of technology, often starting with its use in government itself, where there is, at least in theory, less friction between organizational goals and societal values.

In the United States, New York City has made a public commitment to opening the black box of the government's use of technology: In 2018, an ADS task force was convened, the first of such in the nation, and charged with providing recommendations to New York City's government agencies for how to become transparent and accountable in their use of ADS. In a 2019 report, the task force recommended using ADS where they are beneficial, reduce potential harm and promote fairness, equity, accountability and transparency[1]. Can these principles become policy in the face of the apparent lack of trust in the government's ability to manage AI in the interest of the public? We argue that overcoming this mistrust hinges on our ability to engage in substantive multi-stakeholder conversations around ADS, bringing with it the imperative of interpretability — allowing humans to *understand and, if necessary, contest* the computational process and its outcomes.

Remarkably little is known about how humans perceive and evaluate algorithms and their outputs, what makes a human trust or mistrust an algorithm[2], and how we can empower humans to exercise agency — to adopt or challenge an algorithm's decision. Consider, for example, scoring and ranking — data-driven algorithms that prioritize entities such as individuals, schools, or products and services. These algorithms may be used to determine credit worthiness, and desirability for college admissions or employment. Scoring and ranking are as ubiquitous and powerful as they are opaque. Despite their importance, members of the public often know little about why one person is ranked higher than another by a résumé screening or a credit scoring tool, how the ranking process is designed and whether its results can be trusted.

As an interdisciplinary team of scientists in computer science and social psychology, we propose a *framework* that fuses connections between interpretability and trust, and develops actionable explanations for a diversity of stakeholders, recognizing their unique perspectives and needs. We focus on three questions (Box 1) about making machines interpretable: (1) what are we explaining, (2) to whom are we explaining and for what purpose, and (3) how do we know that an explanation is effective? By asking — and charting the path towards answering — these questions, we can promote greater trust in algorithms, and improve fairness and efficiency of algorithm-assisted decision making.

Box 1 | Research questions

- **What are we explaining?** Do people trust algorithms more or less than they would trust an individual making the same decisions? What are the perceived trade-offs between data disclosure and the privacy of individuals whose data are being analyzed, in the context of interpretability? Which potential sources of bias are most likely to trigger distrust in algorithms? What is the relationship between the perceptions about a dataset's fitness for use and the overall trust in the algorithmic system?
- **To whom are we explaining and why?** How do group identities shape perceptions about algorithms? Do people lose trust in algorithmic decisions when they learn that outcomes produce disparities? Is this only the case when those disparities harm their in-group? Are people more likely to see algorithms as biased if members of their own group were not involved in algorithm construction? What kinds of transparency will promote trust, and when will transparency decrease trust? Do people trust the moral cognition *embedded* within algorithms? Does this apply to some domains (for example, pragmatic decisions, such as clothes shopping) more than others (for example, moral domains, such as criminal sentencing)? Are certain sociodemographics to delegate to algorithms (for example, religious advice)?
- **Are explanations effective?** Do people understand the label? What kinds of explanations allow individuals to exercise agency: make informed decisions, modify their behaviour in light of the information, or challenge the results of the algorithmic process? Does the nutrition label help create trust? Can the creation of nutrition labels lead programmers to alter the algorithm?

What are we explaining?

Existing legal and regulatory frameworks, such as the US's Fair Credit Reporting Act and the EU's General Data Protection Regulation, differentiate between two kinds of explanations. The first concerns the outcome: what are the results for an individual, a demographic group or the population as a whole? The *second* concerns the logic behind the decision-making process: what features help an individual or group get a higher score, or, more generally, what are the rules by which the score is computed? Selbst and Barocas argue for an additional kind of an explanation that considers the justification: why are the rules what they are? Much has been written about explaining outcomes[3], an area focused on explaining and justifying the process.

Procedural justice aims to ensure that algorithms are perceived as fair and

Nutritional Labels for Data and Models *

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

Abstract

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often used outside of the original context for which they were intended. In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a nutritional label, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Nutritional labels are derived automatically or semi-automatically as part of the complex process that gave rise to the data or model they describe, embodying the paradigm of interpretability-by-design. In this paper we further motivate nutritional labels, describe our instantiation of this paradigm for algorithmic rankers, and give a vision for developing nutritional labels that are appropriate for different contexts and stakeholders.

1 Introduction

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often repurposed — used outside of the original context for which they were intended.[1] In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a *nutritional label*, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Short of setting up a chemistry lab, the consumer would otherwise

Copyright 2019 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work was supported in part by NSF Grants No. 1926250, 1916647, and 1741evaluation.

[1] *See Section 3.4 of Salganik's "Bit by Bit" [24] for a discussion of data repurposing in the Digital Age, which he aptly describes as "mixing marzipan with cottonseeds."*

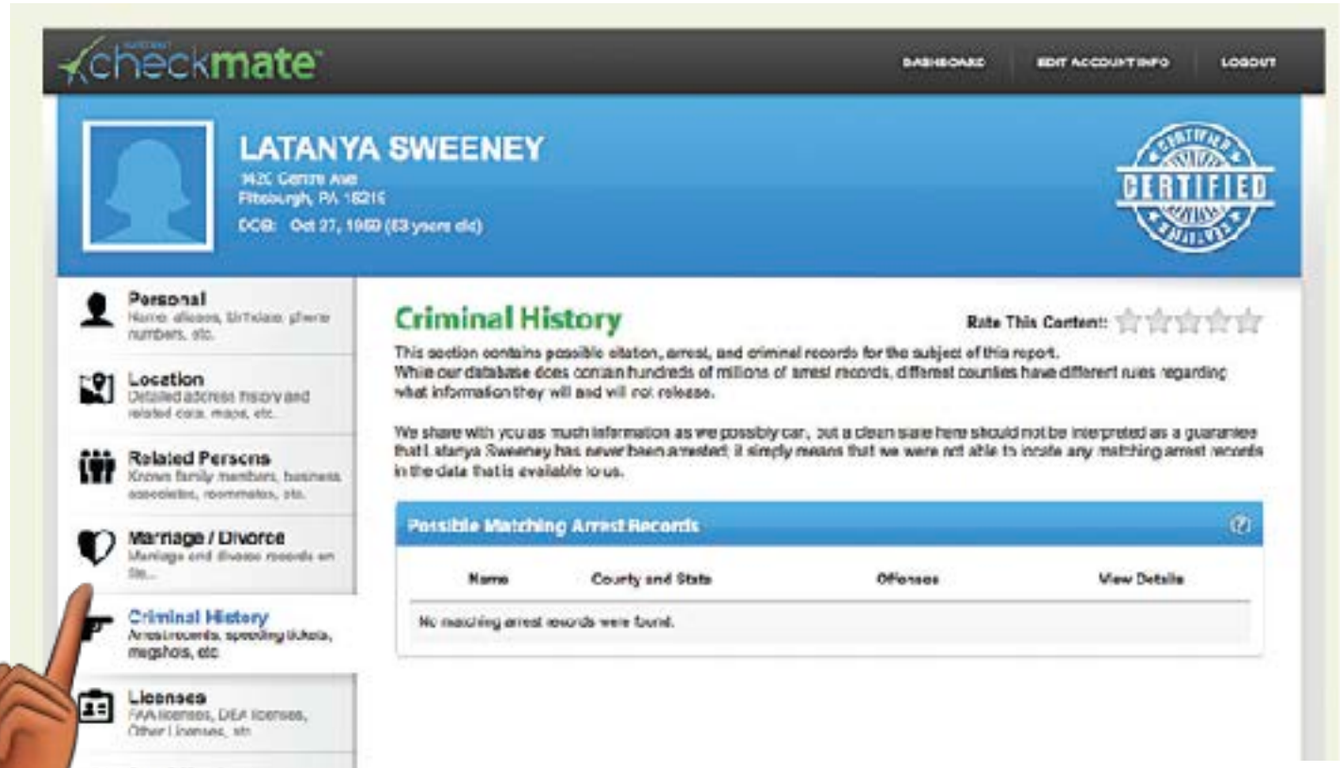
r/ai

Recall: discrimination in online ad delivery

r/ai

Instant Checkmate

Google
AdSense



February 2013

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/

r/ai

Possible explanations

February 2013

Conjectures

Does **Instant Checkmate** **serve ads** specifically for Black-identifying names?

Is **Google** AdSense **explicitly biased** in this way?

Does **Google** AdSense **learn racial bias from click-through rates**?

https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/

Response

Google:"AdWords does not conduct any racial profiling. …It is up to individual advertisers to decide which keywords they want to choose to trigger their ads."

"**Instant Checkmate** would like to state unequivocally that it has never engaged in racial profiling in Google AdWords. We have absolutely no technology in place to even connect a name with a race and have never made any attempt to do so."

r/ai

AdFisher

theguardian

July 2015

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs

One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

Women less likely to be shown ads for high-paid jobs on Google, study shows

The AdFisher tool simulated job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender.

One experiment showed that Google displayed ads for a career coaching service for "$200k+" executive jobs **1,852 times to the male group and only 318 times to the female group**. Another experiment, in July 2014, showed a similar trend but was not statistically significant.

https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

r/ai

AdFisher: Who is responsible?

Finding

(a) (b)

Figure 1: Ads approved by Google in 2015. The ad in the left (right) column was targeted to women (men).

Conjectures

Is **Google** **explicitly programming** the system to show the ad less often to women?

Is the **advertiser** targeting the ad through **explicit use of demographic categories or selection of proxies**, and Google respecting these targeting criteria?

Are **other advertisers** **outbidding** our advertiser when targeting to women?

Are **male and female users** **behaving differently** in response to ads?

[Datta, Datta, Makagon, Mulligan & Tschantz, 2018]

r/ai

Discrimination through optimization

Key question: does the platform itself introduce demographic skew in ad delivery?

Conjectures

Users see relevant ads, maximizing the likelihood of engagement. Based on historical engagement data, delivery may be skewed in ways that an advertiser may not have intended.

Market effects and financial optimization can lead to skewed ad delivery. In a nutshell: some populations are more “valuable” and so advertising to them costs more. If an advertiser bids less, they won’t get to the more “valuable” population.

[Ali, Sapiezynski, Bogen, Korolova, Mislove & Rieke, 2019]

r/ai

Discrimination through optimization

Key question: does the platform itself introduce demographic skew in ad delivery?

Findings

Skew can arise due to financial optimization effects and the ad delivery platform's predictions about the relevance of its ads to different user categories

Ad content - text and images - and advertiser budget both may contribute to the skew.

[Ali, Sapiezynski, Bogen, Korolova, Mislove & Rieke, 2019]

r/ai

Discrimination through optimization

Key question: does the platform itself introduce demographic skew in ad delivery?

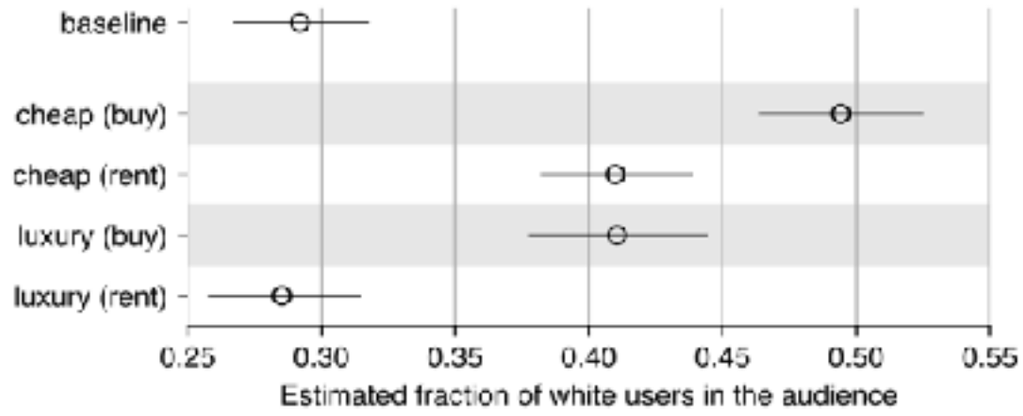


Figure 9: Results for housing ads, showing a breakdown in the ad delivery audience by race. Despite being targeted in the same manner, using the same bidding strategy, and being run at the same time, we observe significant skew in the makeup of the audience to whom the ad is delivered (ranging from estimated 27% white users for luxury rental ads to 49% for cheap house purchase ads).

Findings

Skew was observed along racial lines, in ads for housing opportunities

[Ali, Sapiezynski, Bogen, Korolova, Mislove & Rieke, 2019]

r/ai

legal implications

r/ai

Conclusions from AdFisher

- Each actor in the advertising ecosystem may have contributed inputs that produced the effect
- **It is impossible to know, without additional information, what the different actors - other than the consumers of the ads - did or did not do**
- In particular, impossible to asses intent, which *may* be necessary to asses the extent of legal liability. Or it may not!
  - **Title VII of the 1964 Civil Rights Act** makes it unlawful to discriminate based on sex in several stages of employment. It includes an **advertising prohibition** (think sex-specific *help wanted* columns in a newspaper), which does not turn on intent
  - **Title VII does not directly apply here** because it is limited in scope to employers, labor organizations, employment agencies, joint labor-management committees
  - **Fair Housing Act (FHA)** is perhaps a better guide than Title VII, limiting both content and activities that target advertisement based on protected attributes

[A. Datta, A. Datta, J. Makagon, D. Mulligan, M. Tschantz, 2018]

r/ai

Facebook ads and the Fair Housing Act

THE VERGE

Facebook has been charged with housing discrimination by the US government

'Facebook is discriminating against people based upon who they are and whe they live,' says HUD secretary

By Russell Brandom | Mar 28, 2019, 7:51am EDT

The Department of Housing and Urban Development has filed charges ag for housing discrimination, escalating the company's ongoing fight over di ad targeting system. The charges build on a complaint filed in August, fin reasonable cause to believe Facebook has served ads that violate the Fair Housing Act.

ProPublica first raised concerns over housing discrimination on Facebook in 2016, when reporters found that the "ethnic affinities" tool could be used to exclude black or Hispanic users from seeing specific ads. If those ads were for housing or employment opportunities, the targeting could easily violate federal law. At the time, Facebook had no internal safeguards in place to prevent such targeting.

Fair Housing Act, also called **Title VIII of the Civil Rights Act of 1968**, U.S. federal legislation that protects individuals and families from discrimination in the sale, rental, financing, or **advertising** of housing. The Fair Housing Act, as amended in 1988, prohibits discrimination on the basis of **race**, **color**, **religion**, **sex**, **disability**, **family status**, and **national origin**.

https://www.theverge.com/2019/3/28/18285178/facebook-hud-lawsuit-fair-housing-discrimination

r/ai

The Fair Housing Act

THE VERGE

Facebook has been charged with housing discrimination by the US government

'Facebook is discriminating against people based upon who they are and where they live,' says HUD secretary

By Russell Brandom | Mar 28, 2019, 7:51am EDT

Facebook has struggled to effectively address the possibility of discriminatory ad targeting. The company pledged to step up anti-discrimination enforcement in the wake of *ProPublica*'s reporting, but a follow-up report in 2017 found the same problems persisted nearly a year later.

"WE'RE DISAPPOINTED BY TODAY'S DEVELOPMENTS," FACEBOOK SAYS

According to the HUD complaint, many of the options for targeting or excluding audiences are shockingly direct, including a map tool that explicitly echoes redlining practices. "[Facebook] has provided a toggle button that enables advertisers to exclude men or women from seeing an ad, a search-box to exclude people who do not speak a specific language from seeing an ad, and a map tool to exclude people who live in a specified area from seeing an ad by drawing a red line around that area," the complaint reads.

https://www.theverge.com/2019/3/28/18285178/facebook-hud-lawsuit-fair-housing-discrimination

[Ali, Sapiezynski, Bogen, Korolova, Mislove & Rieke, 2019]

r/ai

Facebook ads and HEC

When is “skew” in fact discrimination?

SUMMARY OF SETTLEMENTS BETWEEN CIVIL RIGHTS ADVOCATES AND FACEBOOK

Housing, Employment, and Credit Advertising Reforms

In the settlements, Facebook will undertake far-reaching changes and steps that will prevent discrimination in housing, employment, and credit advertising on Facebook, Instagram, and Messenger. These changes demonstrate real progress.

- **Facebook will establish a separate advertising portal for creating housing, employment, and credit (“HEC”) ads on Facebook, Instagram, and Messenger that will have limited targeting options, to prevent discrimination.**
- **The following rules will apply to creating HEC ads.**
  - *Gender, age, and multicultural affinity targeting options will not be available when creating Facebook ads.*
  - *HEC ads must have a minimum geographic radius of 15 miles from a specific address or from the center of a city. Targeting by zip code will not be permitted.*

March 19, 2019

r/ai

Facebook ads and HEC

SUMMARY OF SETTLEMENTS BETWEEN CIVIL RIGHTS ADVOCATES

AND FACEBOOK

Housing, Employment, and Credit Advertising Reforms

- *HEC ads will not have targeting options that describe or appear to be related to personal characteristics or classes protected under anti-discrimination laws. This means that targeting options that may relate to race, color, national origin, ethnicity, gender, age, religion, family status, disability, and sexual orientation, among other protected characteristics or classes, will not be permitted on the HEC portal.*
- *Facebook's "Lookalike Audience" tool, which helps advertisers identify Facebook users who are similar to advertisers' current customers or marketing lists, will no longer consider gender, age, religious views, zip codes, Facebook Group membership, or other similar categories when creating customized audiences for HEC ads.*

r/ai

Facebook ads and HEC

SUMMARY OF SETTLEMENTS BETWEEN CIVIL RIGHTS ADVOCATES

AND FACEBOOK

Housing, Employment, and Credit Advertising Reforms

- *Advertisers will be asked to create their HEC ads in the HEC portal, and if Facebook detects that an advertiser has tried to create an HEC ad outside of the HEC portal, Facebook will block and re-route the advertiser to the HEC portal with limited options.*

r/ai

And it’s not just Facebook

POLICY \ US & WORLD \ TECH \

HUD reportedly also investigating Google and Twitter in housing discrimination probe

By Adi Robertson @thedextriarchy Mar 28, 2019, 3:52pm EDT

POLICY \ US & WORLD \ TECH \

Facebook has been charged with housing discrimination by the US government

'Facebook is discriminating against people based upon who they are and where they live,' says HUD secretary

By Russell Brandom Mar 28, 2019, 7:51am EDT

Fair Housing Act, also called **Title VIII of the Civil Rights Act of 1968**, U.S. federal legislation that protects individuals and families from discrimination in the sale, rental, financing, or **advertising** of housing. The Fair Housing Act, as amended in 1988, prohibits discrimination on the basis of **race**, **color**, **religion**, **sex**, **disability**, **family status**, and **national origin**.

This is the first federal discrimination lawsuit to deal with **racial bias in targeted advertising**, a milestone that lawyers at HUD said was overdue. “Even as we confront new technologies, the fair housing laws enacted over half a century ago remain clear—discrimination in housing-related advertising is against the law,” said HUD General Counsel Paul Compton. “**Just because a process to deliver advertising is opaque and complex doesn’t mean that it’s exempts Facebook and others from our scrutiny and the law of the land**.”

https://www.theverge.com/2019/3/28/18285178/facebook-hud-lawsuit-fair-housing-discrimination

r/ai

The socio-legal landscape

Related concern:
Are ads commercial free speech?

- The First Amendment of the U.S. Constitution protects advertising, but the U.S. Supreme Court set out a test for assessing restrictions on commercial speech, which begins by determining whether the speech is misleading
- Are online ads suggesting the existence of an arrest record misleading if no one by that name has an arrest record?
- Assume the ads are free speech: what happens when these ads appear more often for one racial group than another? Not everyone is being equally affected. Is that free speech or racial discrimination?

[Sweeney, 2013]

r/ai

Tracking and consent

The New York Times

To Be Tracked or Not? Apple Is Now Giving Us the Choice.

By Brian X. Chen

April 26, 2021 Updated 12:40 p.m. ET

If we had a choice, would any of us want to be tracked online for the sake of seeing more relevant digital ads?

We are about to find out.

On Monday, Apple plans to release iOS 14.5, one of its most anticipated software updates for iPhones and iPads in years. It includes a new privacy tool, called App Tracking Transparency, which could give us more control over how our data is shared.

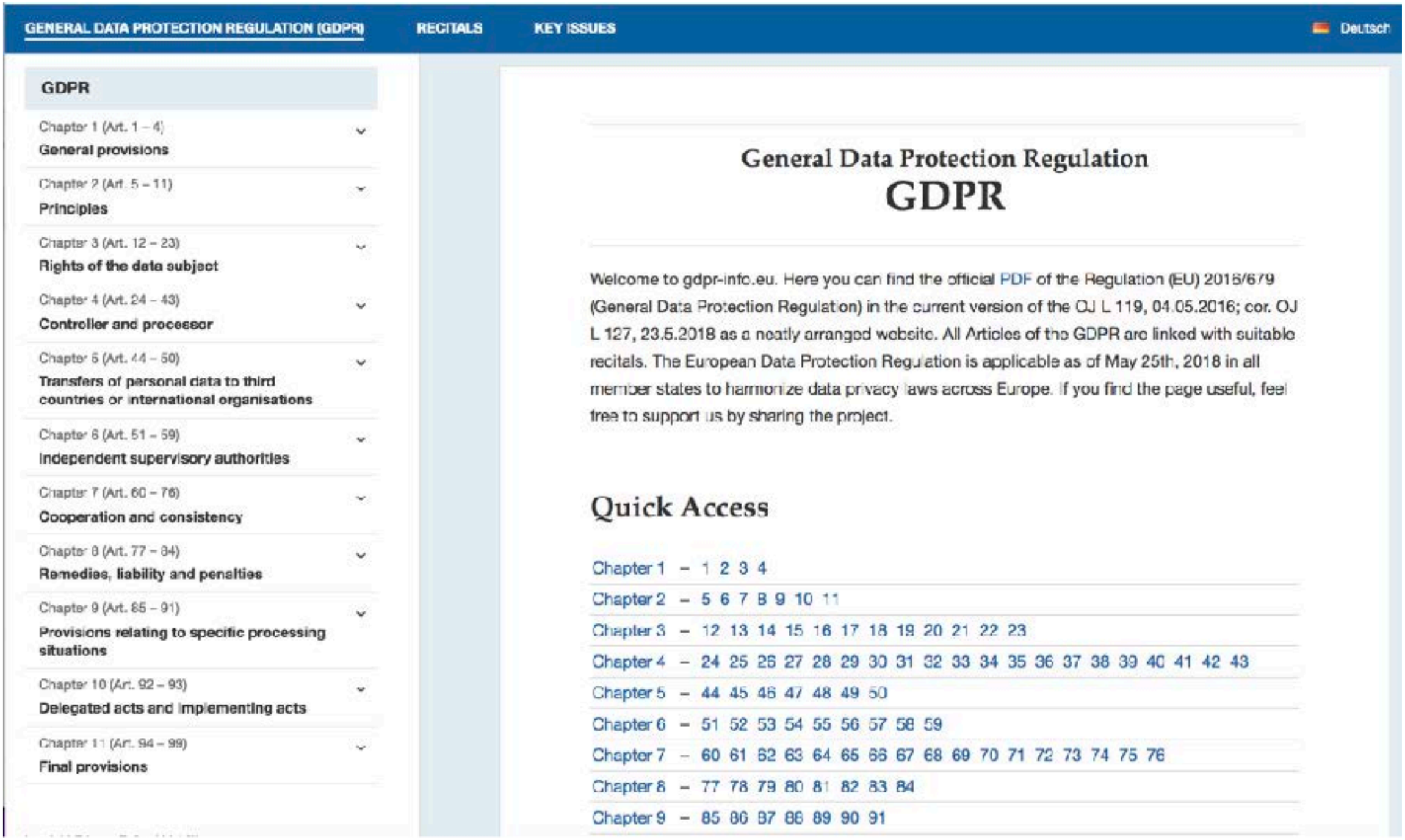
Here's how it works: When an app wants to follow our activities to share information with third parties such as advertisers, a window will show up on our Apple device to ask for our permission to do so. If we say no, the app must stop monitoring and sharing our data.

A pop-up window may sound like a minor design tweak, but it has thrown the online advertising industry into upheaval. Most notably, Facebook has gone on the warpath. Last year, the social network created a website and took out full-page ads in newspapers denouncing Apple's privacy feature as harmful to small businesses.

r/ai

data protection:
the GDPR

r/ai



adopted in April 2016

enforced since May 25, 2018

https://gdpr-info.eu/

r/ai

GDPR: scope and definitions

Article 2: Material Scope

- This Regulation applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.

Article 4: Definitions

- ‘**personal data**’ means any information relating to an identified or identifiable natural person (‘**data subject**’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

- ‘**processing**’ means **any operation** or set of operations which is performed on personal data or on sets of personal data, **whether or not by automated means**, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;

GDPR: scope and definitions

Article 4: Definitions

- ‘**controller**’ means the natural or legal person, public authority, agency or other body which, alone or jointly with others, **determines the purposes and means of the processing** of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;
- ‘**processor**’ means a natural or legal person, public authority, agency or other body which **processes personal data on behalf of the controller**;
- ‘**consent**’ of the data subject means any **freely given, specific, informed and unambiguous** indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, **signifies agreement to the processing of personal data** relating to him or her;

Art. 7 GDPR

Conditions for consent

1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.

2. [1] If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. [2] Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.

Art. 7 GDPR

Conditions for consent

3. [1] The data subject shall have the right to withdraw his or her consent at any time. [2] The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. [3] Prior to giving consent, the data subject shall be informed thereof. [4] It shall be as easy to withdraw as to give consent.

4. When assessing whether consent is freely given, utmost account shall be taken of whether, *inter alia*, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.

Chapter 3

Rights of the data subject

Section 1	–	**Transparency and modalities**
Article 12	–	Transparent information, communication and modalities for the exercise of the rights of the data subject
Section 2	–	**Information and access to personal data**
Article 13	–	Information to be provided where personal data are collected from the data subject
Article 14	–	Information to be provided where personal data have not been obtained from the data subject
Article 15	–	Right of access by the data subject

Chapter 3

Rights of the data subject

Section 3	–	**Rectification and erasure**
Article 16	–	Right to rectification
Article 17	–	Right to erasure ('right to be forgotten')
Article 18	–	Right to restriction of processing
Article 19	–	Notification obligation regarding rectification or erasure of personal data or restriction of processing
Article 20	–	Right to data portability

Removing personal data

The right to be forgotten (Article 17)

- Similar laws exist in other jurisdictions, e.g., Argentina (since 2006)
- Resulted in many dereferencing requests to search engines
- Often seen as controversial: **reasons?**
- May conflict with other legal requirements, or with technical requirements

- **Also, just technically challenging:**
  - have to re-engineer the data management stack, **what are the issues**?
  - what about models?

[S. Abiteboul and J. Stoyanovich, 2019]

r/ai

Chapter 3

Rights of the data subject

Section 3	–	**Rectification and erasure**
Article 16	–	Right to rectification
Article 17	–	Right to erasure ('right to be forgotten')
Article 18	–	Right to restriction of processing
Article 19	–	Notification obligation regarding rectification or erasure of personal data or restriction of processing
Article 20	–	Right to data portability

Moving personal data

The right to data portability (Article 20)

- Aims to prevent vendor lock-in
- What are some technical difficulties?
  - Suppose you want to move your photos from Service A to Service B?
  - What about moving your social interactions from Service A to Service B?
- Can we look at this from the point of view of **inter-operability** rather than moving data?

[S. Abiteboul and J. Stoyanovich, 2019]

r/ai

Moving personal data

Data Transfer Project

Download White Paper

About Community Documentation Updates FAQ

About us

The Data Transfer Project was launched in 2018 to create an open-source, service-to-service data portability platform so that all individuals across the web could easily move their data between online service providers whenever they want.

The contributors to the Data Transfer Project believe portability and interoperability are central to innovation. Making it easier for individuals to choose among services facilitates competition, empowers individuals to try new services and enables them to choose the offering that best suits their needs.

Current contributors include:

What is the Data Transfer Project

Data Transfer Project (DTP) is a collaboration of organizations committed to building a common framework with open-source code that can connect any two online service providers, enabling a seamless, direct, user initiated portability of data between the two platforms.

Learn More

https://datatransferproject.dev/

r/ai

Chapter 3

Rights of the data subject

Section 4 – Right to object and automated individual decision-making

Article 21 – Right to object

Article 22 – Automated individual decision-making, including profiling

Recital 58

The principle of transparency*

[1] The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used. [2] Such information could be provided in electronic form, for example, when addressed to the public, through a website. [3] This is of particular relevance in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected, such as in the case of online advertising. [4] Given that children merit specific protection, any information and communication, where processing is addressed to a child, should be in such a clear and plain language that the child can easily understand.

from data to impacts:
algorithmic impact
statements

r/ai

Regulating ADS?

Precautionary

@FalaahArifKhan

Nah! I'm fine!

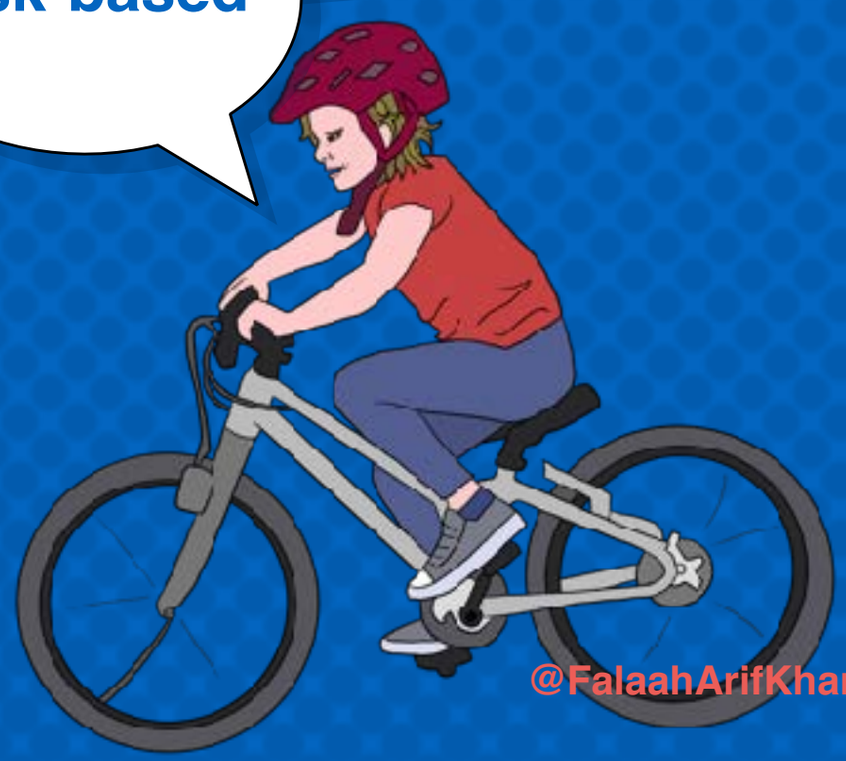
@FalaahArifKhan

The Anti-Elon
@antiElon

Regulation rocks!

2.3K 9.2K 126K

Risk-based



@FalaahArifKhan

Setting the stage: “Big Data Policing”

“Despite its growing popularity, predictive policing is in its relative infancy and is still mostly hype. Current prediction is akin to early weather forecasting, and, like Big Data approaches in other sectors, mixed evidence exists about its effectiveness.

Cities such as Los Angeles, Atlanta, Santa Cruz, and Seattle have enlisted the predictive policing software company PredPol to predict where property crimes will occur. Santa Cruz reportedly “saw burglaries drop by 11% and robberies by 27% in the first year of using [PredPol’s] software.” Similarly, Chicago’s Strategic Subject List—or “heat list”—of people most likely to be involved in a shooting had, as of mid-2016, predicted more than 70% of the people shot in the city, according to the police.

But two rigorous academic evaluations of predictive policing experiments, one in Chicago and another in Shreveport, have shown no benefit over traditional policing. **A great deal more study is required to measure both predictive policing’s benefits and its downsides**. “

what are the potential benefits?

what are the potential downsides?

[Andrew Selbst, 2017]

r/ai

How to regulate “Big Data Policing”

“While policing is just one of many aspects of society being upended by machine learning, and potentially exacerbating disparate impact in a hidden way as a result, it is a particularly useful case study because of how little our legal system is set up to regulate it.”

The Fourth Amendment: *The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.*

“[…] the Fourth Amendment’s reasonable suspicion requirement is inherently a “small data doctrine,” rendering it impotent in even its primary uses when it comes to data mining.”

new legal strategies are needed

[Andrew Selbst, 2017]

r/ai

How to regulate “Big Data Policing”

“ Regarding predictive policing specifically, society lacks basic knowledge and transparency about both the technology’s efficacy and its effects on vulnerable populations. Thus, this Article proposes a regulatory solution designed to fill this knowledge gap—to make the police do their homework and show it to the public before buying or building these technologies.”

Main contribution: Algorithmic Impact Statements (AISs)

“Impact statements are designed to **force consideration of the problem at an early stage**, and to document the process so that the public can learn what is at stake, **perhaps as a precursor to further regulation**. The primary problem is that no one, including the police using the technology, yet knows what the results of its use actually are.”

[Andrew Selbst, 2017]

r/ai

Algorithmic Impact Statements (AISs)

- Modeled on the Environmental Impact Statements (EISs) of the 1969 National Environmental Policy Act (NEPA)
- GDPR requires “data protection impact assessments (DPIAs) whenever data processing “is likely to result in a high risk to the rights and freedoms of natural persons”
- Privacy impact statements (PIAs) are used to assess the risks of using personally identifiable information by IT systems

The gist:

- Explore and evaluate all reasonable alternatives
- Include the alternative of “No Action”
- Include appropriate mitigation measures
- Provide opportunities for public comment

[Andrew Selbst, 2017]

r/ai

Canadian ADS directive

r/ai

Government of Canada | Gouvernement du Canada

Home → How government works → Policies, directives, standards and guidelines

Directive on Automated Decision-Making

The Government of Canada is increasingly looking to utilize artificial intelligence to make, or assist in making, administrative decisions to improve service delivery. The Government is committed to doing so in a manner that is compatible with core administrative law principles such as transparency, accountability, legality, and procedural fairness. Understanding that this technology is changing rapidly, this Directive will continue to evolve to ensure that it remains relevant.

Date modified: 2019-02-05

- Took effect on April 1, 2019, compliance by April 1, 2020
- Applies to any ADS developed or procured after April 1, 2020
- **Reviewed automatically every 6 months**

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

Definitions

Appendix A: Definitions

- **Administrative Decision** Any decision that is made by an authorized official of an institution as identified in section 9 of this Directive pursuant to powers conferred by an Act of Parliament or an order made pursuant to a prerogative of the Crown that affects legal rights, privileges or interests.
- **Algorithmic Impact Assessment** A framework to help institutions better understand and reduce the risks associated with Automated Decision Systems and to provide the appropriate governance, oversight and reporting/ audit requirements that best match the type of application being designed.
- **Automated Decision System** Includes any technology that either assists or replaces the judgement of human decision-makers. These systems draw from fields like statistics, linguistics, and computer science, and use techniques such as rules-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets.

Objectives

Section 4: Objectives and Expected Results

- **4.1** The objective of this Directive is to ensure that Automated Decision Systems are deployed in a manner that **reduces risks** to Canadians and federal institutions, and **leads to more efficient, accurate, consistent, and interpretable decisions** made pursuant to Canadian law.
- **4.2** The expected results of this Directive are as follows:
  - Decisions made by federal government departments are data-driven, responsible, and complies with procedural fairness and due process requirements.
  - Impacts of algorithms on administrative decisions are assessed and negative outcomes are reduced, when encountered.
  - Data and information on the use of Automated Decision Systems in federal institutions are made available to the public, when appropriate.

Requirements

Section 6.1: Algorithmic Impact Assessment (excerpt)

- **6.1.1 Completing** an Algorithmic Impact Assessment **prior to the production** of any Automated Decision System.
- **6.1.2** …
- **6.1.3 Updating** the Algorithmic Impact Assessment when system functionality or the scope of the Automated Decision System changes.
- **6.1.4 Releasing the final results of Algorithmic Impact Assessments** in an accessible format via Government of Canada websites and any other services designated by the Treasury Board of Canada Secretariat pursuant to the Directive on Open Government.

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

Requirements

Section 6.2: Transparency

- providing notice before decisions
- providing explanations after decisions
- access to components
- release of source code, unless it’s classified Secret, Top Secret or Protected C

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

Impact Assessment Levels

Decisions classified w.r.t. impact on:

- the rights of individuals or communities,
- the health or well-being of individuals or communities,
- the economic interests of individuals, entities, or communities,
- the ongoing sustainability of an ecosystem.

Level I: no impact: impacts are reversible and brief

Level II: moderate: impacts are likely reversible and short-term

Level III: high: impacts are difficult to reversible and ongoing

Level IV: very high: impacts are irreversible and perpetual

higher impact levels lead to more stringent requirements

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

regulating ADS in New York City

r/ai

How it started: The Vacca bill

August 16, 2017

Int. No. 1696

By Council Member Vacca

A Local Law to amend the administrative code of the city of New York, in relation to automated processing of data for the purposes of targeting services, penalties, or policing to persons

Be it enacted by the Council as follows:

Section 1. Section 23-502 of the administrative code of the city of New York is amended to add a new subdivision g to read as follows:

g. Each agency that uses, for the purposes of targeting services to persons, imposing penalties upon persons or policing, an algorithm or any other method of automated processing system of data shall:

1. Publish on such agency's website, the source code of such system; and

2. Permit a user to (i) submit data into such system for self-testing and (ii) receive the results of having such data processed by such system.

§ 2. This local law takes effect 120 days after it becomes law.

MAJ
LS# 10948
8/16/17 2:13 PM

r/ai

How it started: The Vacca bill

October 16, 2017

THE NEW YORKER

By Julia Powles December 20, 2017

ELEMENTS

NEW YORK CITY'S BOLD, FLAWED ATTEMPT TO MAKE ALGORITHMS ACCOUNTABLE

Automated systems guide the allocation of everything from firehouses to food stamps. So why don't we know more about them?

Photograph by Mario Tama / Getty

https://dataresponsibly.github.io/documents/Stoyanovich_VaccaBill.pdf

r/ai

New York City Local Law 49

January 11, 2018

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

Sign In

LEGISLATIVE RESEARCH CENTER

Council Home | Legislation | Calendar | City Council | Committees

RSS | Alerts

Details | Reports

File #:	Int 1696-2017 Version: A	Name:	Automated decision systems used by agencies.
Type:	Introduction	Status:	Enacted
		Committee:	Committee on Technology
On agenda:	8/24/2017		
Enactment date:	1/11/2018	Law number:	2018/049
Title:	A Local Law in relation to automated decision systems used by agencies		
Sponsors:	James Vacca, Helen K. Rosenthal, Corey D. Johnson, Rafael Salamanca, Jr., Vincent J. Gentile, Robert E. Cornegy, Jr., Jumaane D. Williams, Ben Kallos, Carlos Menchaca		
Council Member Sponsors:	9		
Summary:	This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems.		
Indexes:	Oversight		
Attachments:	1. Summary of Int. No. 1696-A, 2. Summary of Int. No. 1696, 3. Int. No. 1696, 4. August 24, 2017 - Stated Meeting Agenda with Links to Files, 5. Committee Report 10/16/17, 6. Hearing Testimony 10/16/17, 7. Hearing Transcript 10/16/17, 8. Proposed Int. No. 1696-A - 12/12/17, 9. Committee Report 12/7/17, 10. Hearing Transcript 12/7/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Transcript - Stated Meeting 12-11-17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 49, 17. Minutes of the Stated Meeting - December 11, 2017		

r/ai

New York City Local Law 49

January 11, 2018

An **Automated Decision System (ADS)** is a "computerized implementation of algorithms, including those derived from machine learning or other data processing or artificial intelligence techniques, which are used to make or assist in making decisions."

Form task force that surveys the current use of ADS in City agencies and develops procedures for:

- requesting and receiving an **explanation** of an algorithmic decision affecting an individual (3(b))
- interrogating ADS for **bias and discrimination** against members of legally-protected groups (3(c) and 3(d))
- allowing the **public** to **assess** how ADS function and are used (3(e)), and archiving ADS together with the data they use (3(f))

r/ai

The ADS task force

May 16, 2018

r/ai

The ADS task force

April 15, 2019

POLICY \ REPORT \ BS & WORLD

New York City's algorithm task force is fracturing

Some members say the city isn't being transparent

By Colin Lecher | @colinlecher | Apr 15, 2019, 8:43am EDT

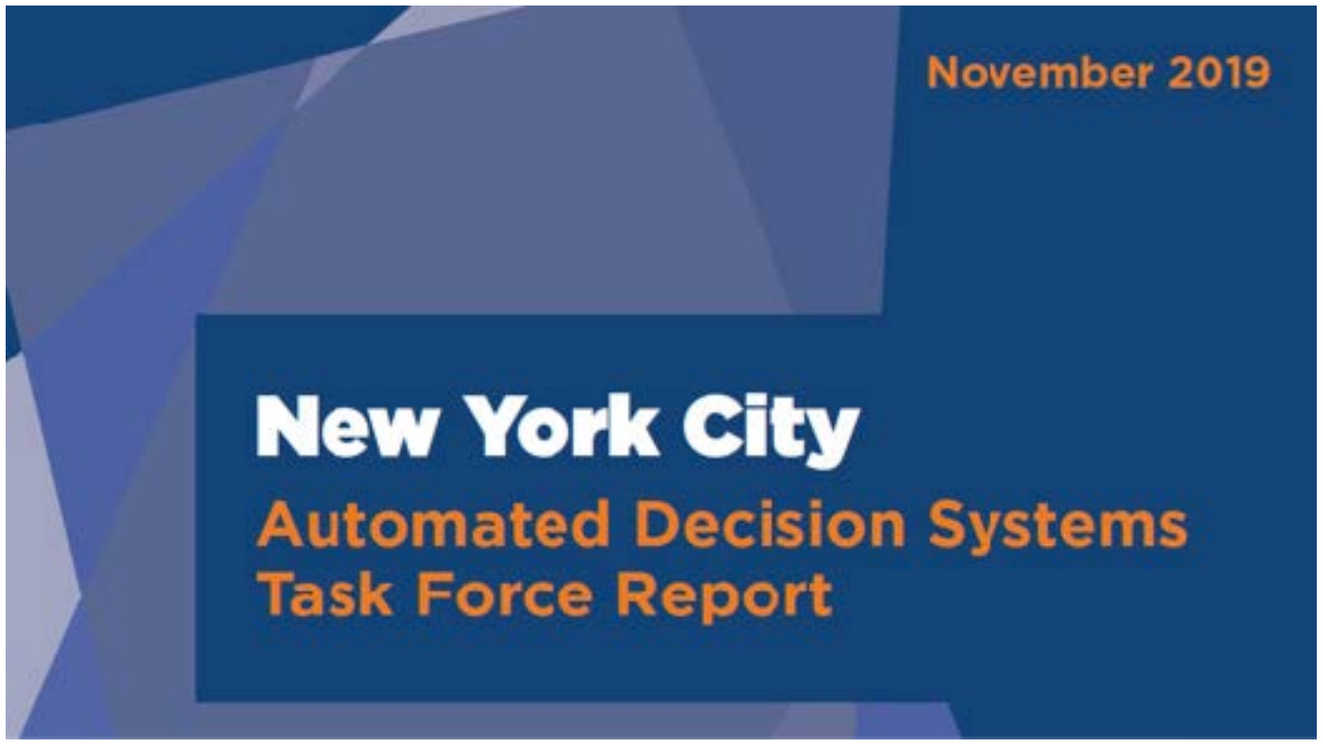
With nothing to study, critics say, the task force is toothless and able to provide only broad policy recommendations …

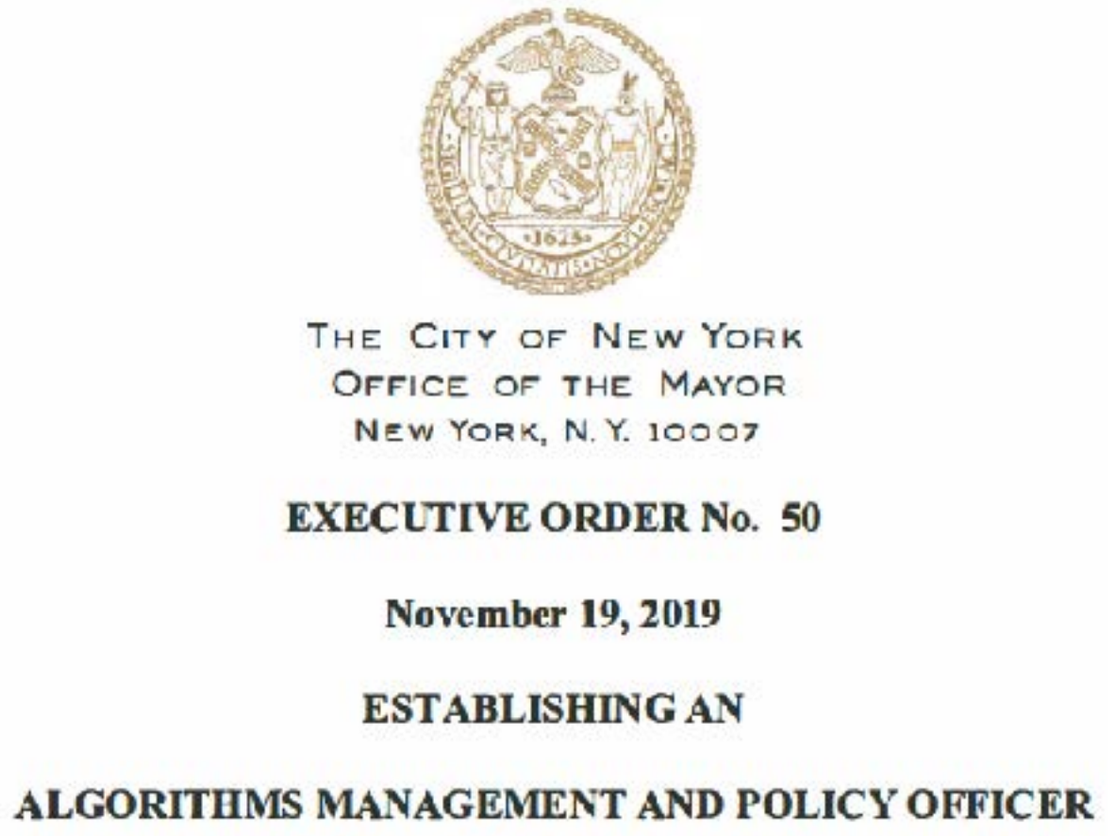
New York University assistant professor and task force member Julia Stoyanovich told *The Verge* that **if no examples are forthcoming, "then there was really no point in forming the task force at all."**

https://dataresponsibly.github.io/documents/StoyanovichBarocas_April4,2019testimony.pdf

The ADS task force

November 19, 2019




THE CITY OF NEW YORK
OFFICE OF THE MAYOR
NEW YORK, N.Y. 10007

EXECUTIVE ORDER No. 50

November 19, 2019

ESTABLISHING AN

ALGORITHMS MANAGEMENT AND POLICY OFFICER

https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf
https://www1.nyc.gov/assets/home/downloads/pdf/executive-orders/2019/eo-50.pdf

r/ai

ADS task force report

@FalaahArifKhan

Principles

- using ADS **where** they promote innovation and efficiency in service delivery
- promoting **fairness**, **equity**, **accountability**, and **transparency** in the use of ADS
- reducing potential harm **across the lifespan** of ADS

Recommendations

- formalize ADS management functions
- build the City's ADS management capacity
- broaden public conversation on ADS

r/ai

so what's algorithmic transparency?

r/ai

Point 1

algorithmic transparency is not synonymous with releasing the source code

publishing source code helps, but it is sometimes unnecessary and often insufficient

Point 2

algorithmic transparency requires data transparency

data is used in training, validation, deployment

validity, accuracy, applicability can only be understood in the data context

data transparency is necessary for all ADS, not only for ML-based systems

Point 3

data transparency is not synonymous with making all data public

release data whenever possible;

also release:

data selection, collection and pre-processing methodologies; data provenance and quality information; known sources of bias; privacy-preserving statistical summaries of the data

Point 3

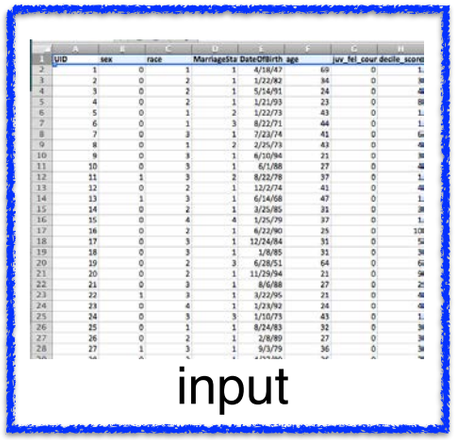
data transparency is not synonymous with making all data public

release data whenever possible;

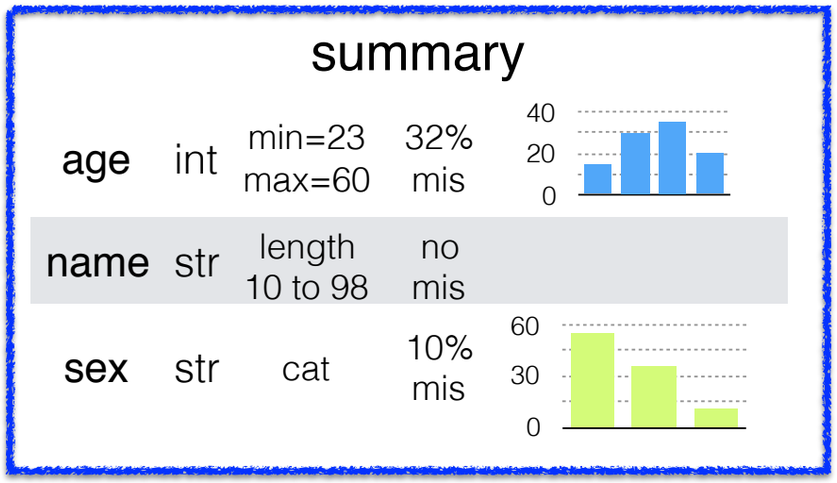
also release:

data selection, collection and pre-processing methodologies; data provenance and quality information; known sources of bias; privacy-preserving statistical summaries of the data

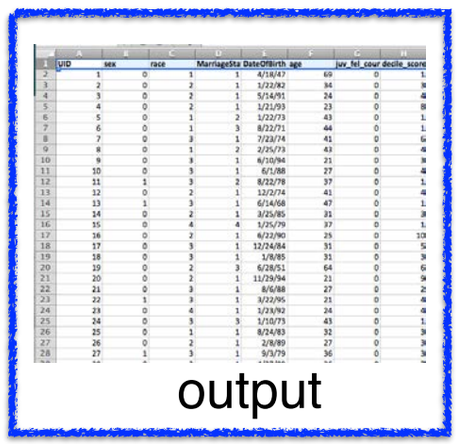
Data Synthesizer



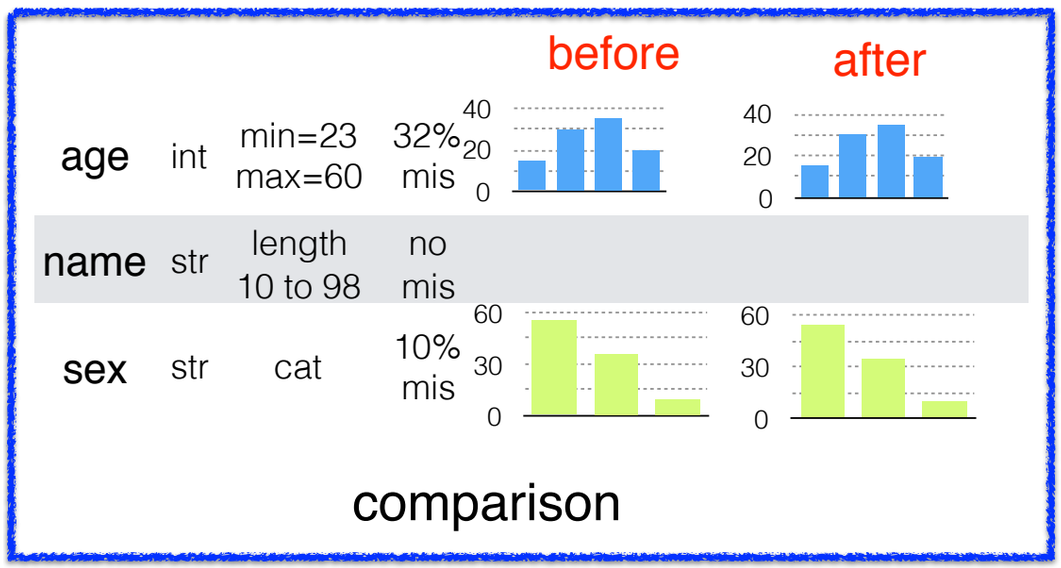
Data Describer



Data Generator



Model Inspector



[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

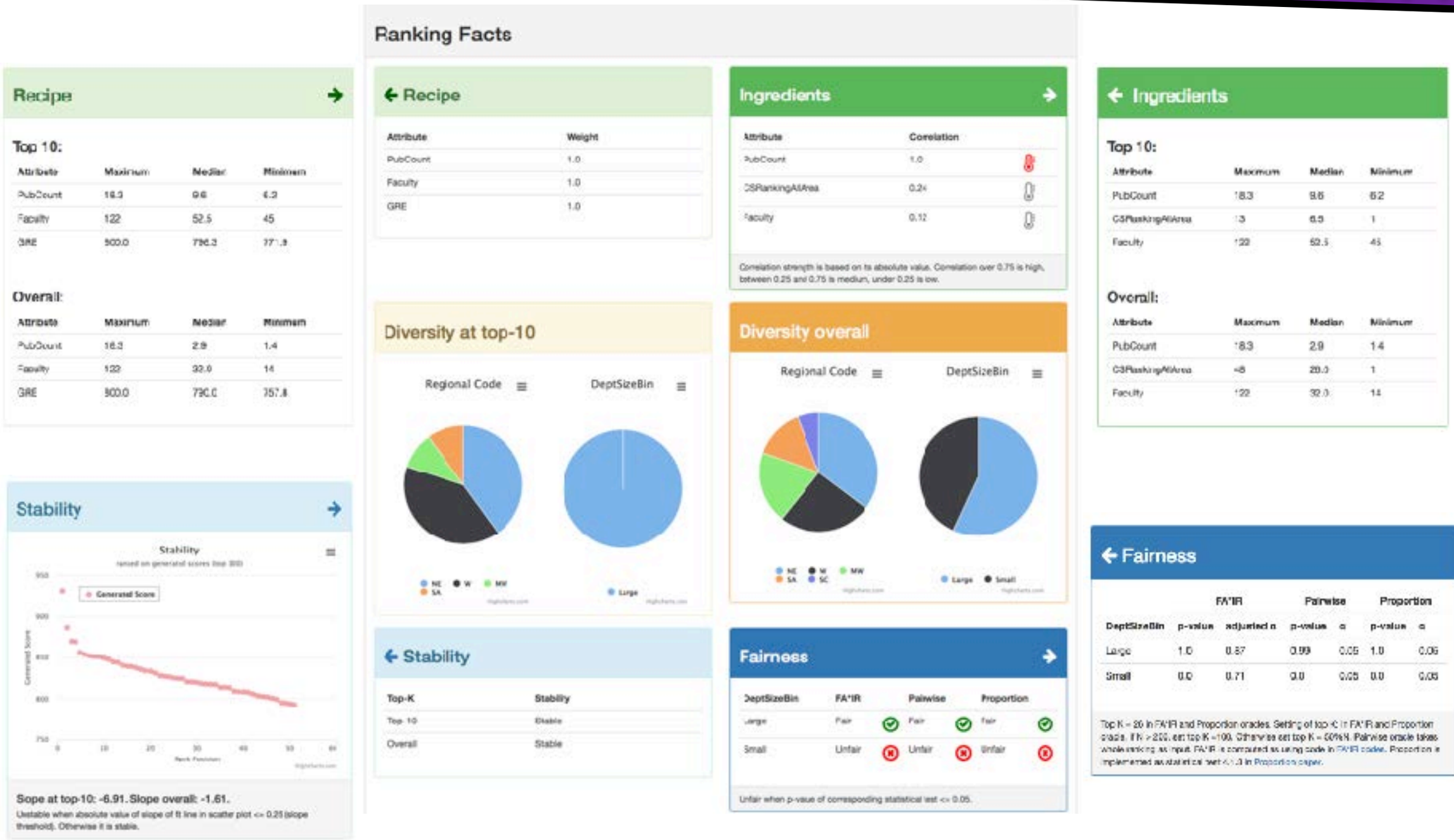
Point 4

actionable transparency requires
interpretability

explain assumptions and effects, not details of operation

engage the public - technical and non-technical

“Nutritional labels” for data and models

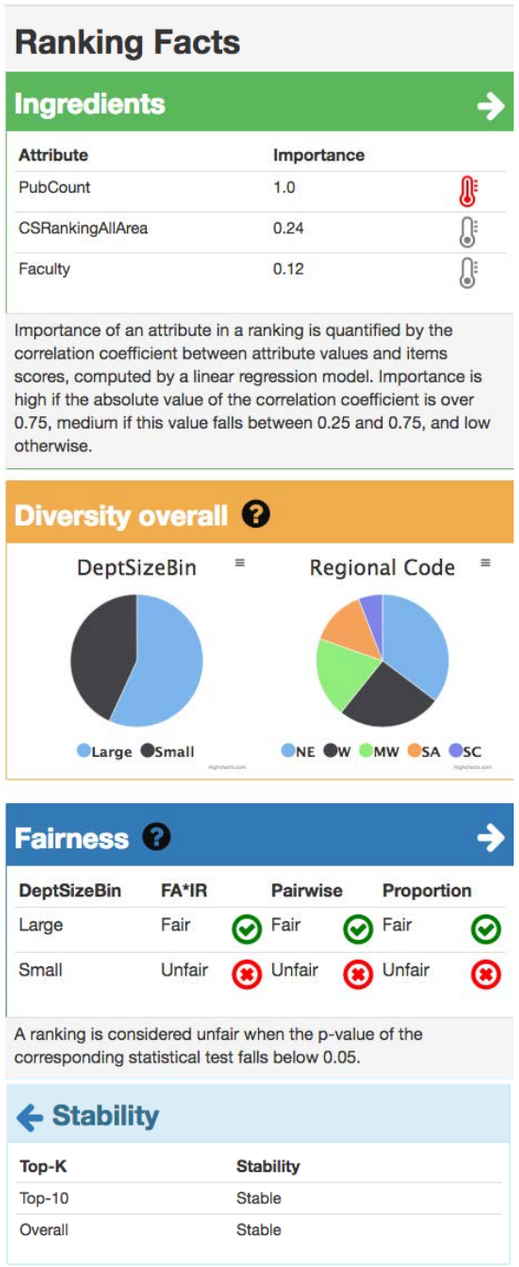


http://demo.dataresponsibly.com/rankingfacts/nutrition_facts/

[K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, HV Jagadish, G. Miklau; 2018]

r/ai

Properties of a nutritional label



comprehensible: short, simple, clear

consultative: provide actionable info

comparable: implying a standard

concrete: helps determine a dataset's fitness for use for a given task

computable: produced as a "by-product" of computation - interpretability-by-design

[Stoyanovich and Howe, 2019]

r/ai

Point 5

transparency / interpretability by design, not as an afterthought

provision for transparency and interpretability at every stage of the data lifecycle

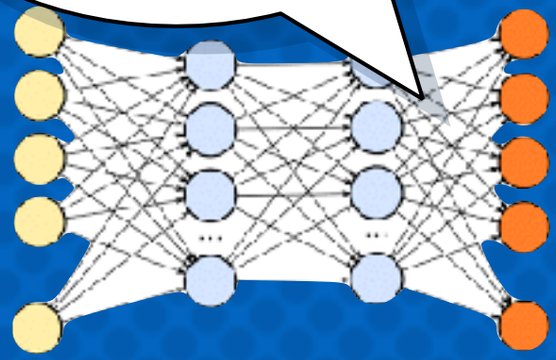
useful internally during development, for communication and coordination between agencies, and for accountability to the public

Frog's eye view

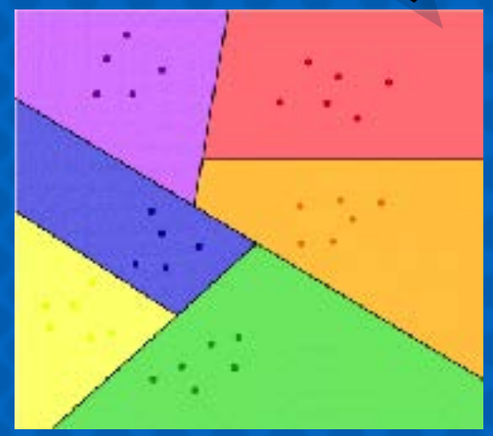
where did the data come from?

	A	B	C	D	E	F	G	H
1	LFD	sex	race	MaritageSta	DateOfBirth	age	juv_tri_cou	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/12/94	21	0	3
11	10	0	3	1	6/5/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/16/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	11/24/84	31	0	3
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/28/94	21	0	9
22	21	0	3	1	8/5/88	27	0	2
23	22	1	3	4	3/22/95	20	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/12/73	43	0	1
26	25	0	1	1	8/26/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/9/79	36	0	4
29	28	0	2	1	4/27/95	20	0	3

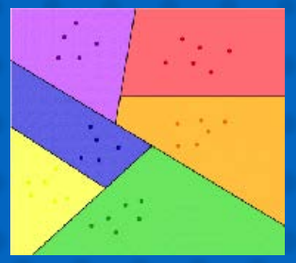
what happens inside the box?



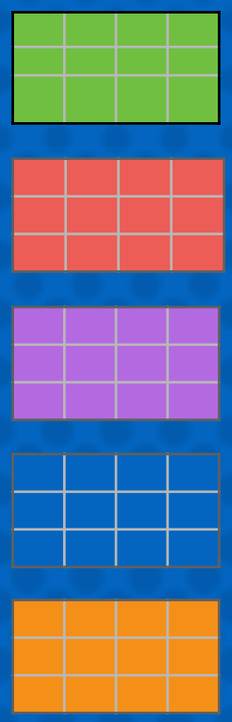
how are results used?



Data lifecycle of an ADS

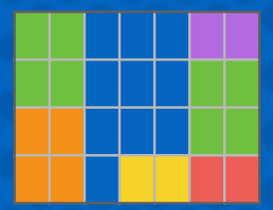


analysis
validation

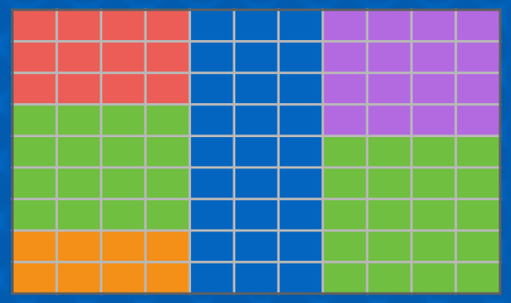


sharing
annotation

querying
ranking



acquisition
curation



interpretability in the eye of the stakeholder

r/ai

What are we explaining?

process (same for everyone? why is this the process?) vs. outcome

procedural justice aims to ensure that algorithms are perceived as fair and legitimate

data transparency is unique to algorithm-assisted decision-making, relates to the justification dimension of interpretability

[J. Stoyanovich, J. Van Bavel, T. West, 2020]

r/ai

To whom are we explaining and why?

accounting for the needs of different stakeholders

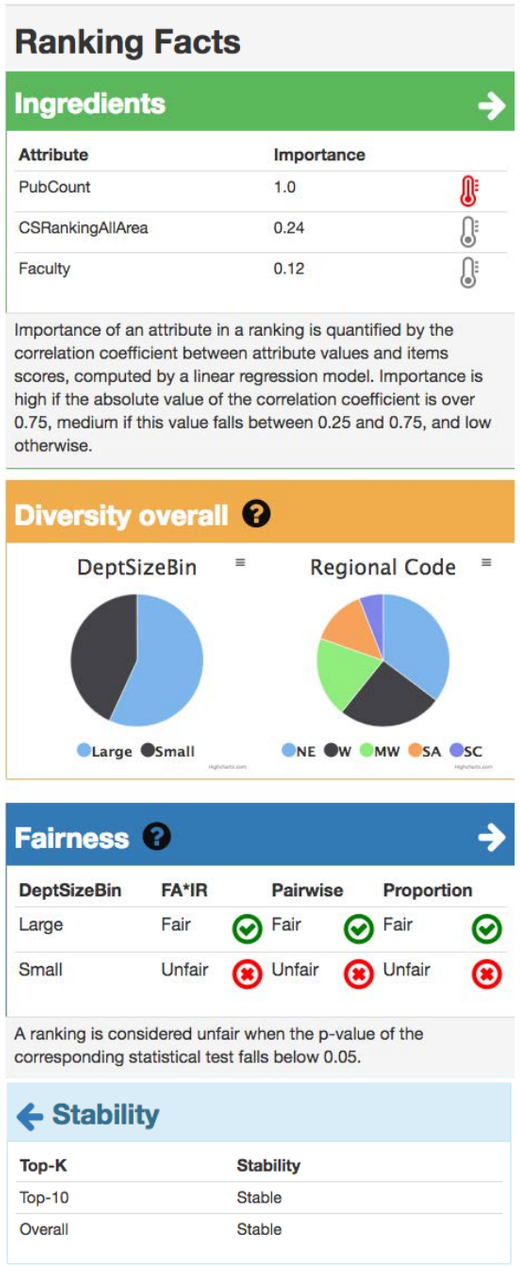
social identity - people trust their in-group members more

moral cognition - is a decision or outcome morally right or wrong?

[J. Stoyanovich, J. Van Bavel, T. West, 2020]

r/ai

How do we know that we explained well?



nutritional labels! :)

… but do they work?

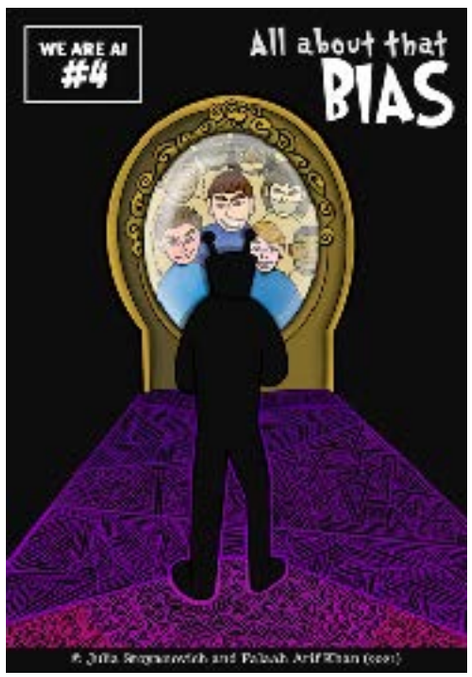
[J. Stoyanovich, J. Van Bavel, T. West, 2020]

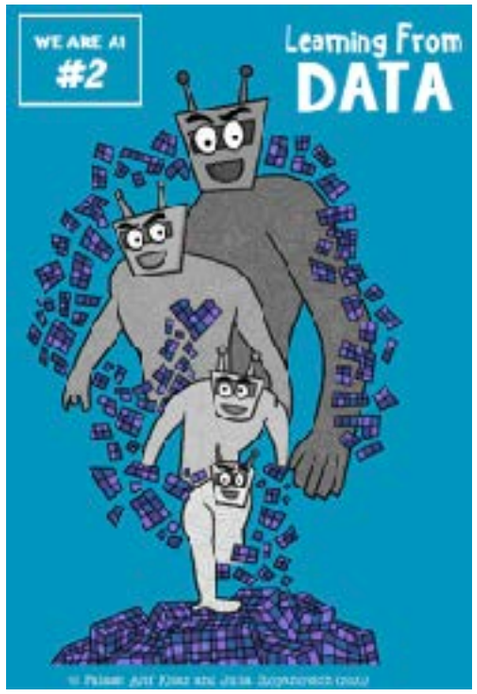
r/ai

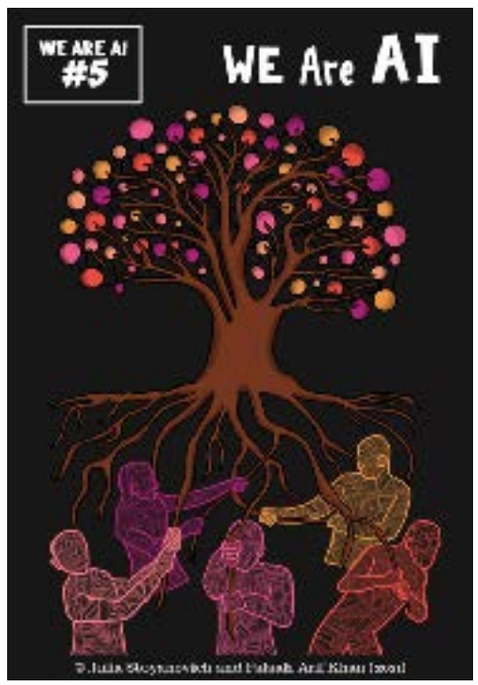
https://dataresponsibly.github.io/we-are-ai/

r/ai

AI comics for the general public







dataresponsibly.github.io/we-are-ai/comics

r/ai

regulating automated hiring systems

r/ai

Regulating hiring ADS: Int 1894-2020

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

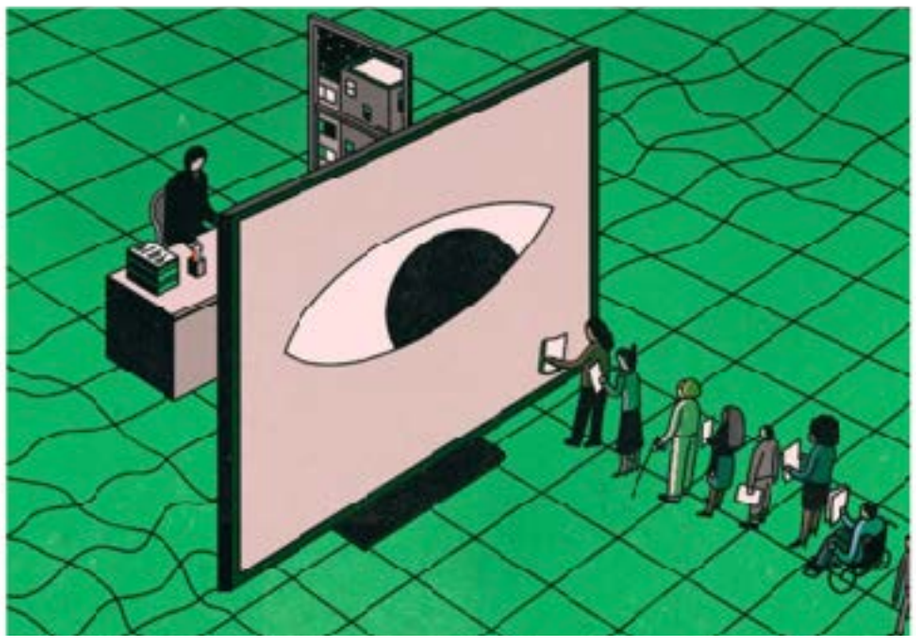
This bill would **regulate the use of automated employment decision tools**, which, for the purposes of this bill, encompass certain systems that use algorithmic methodologies to filter candidates for hire or to make decisions regarding any other term, condition or privilege of employment. This bill would prohibit the sale of such tools if they were not the **subject of an audit for bias** in the past year prior to sale, were not sold with a yearly bias audit service at no additional cost, and were not accompanied by a notice that the tool is subject to the provisions of this bill. This bill would also require any person who uses automated employment assessment tools for hiring and other employment purposes to **disclose to candidates, within 30 days, when such tools were used** to assess their candidacy for employment, and the **job qualifications or characteristics** for which the tool was used to screen. Violations of the provisions of the bill would incur a penalty.

Hiring ADS regulation

The New York Times | March 17, 2021

We Need Laws to Take On Racism and Sexism in Hiring Technology

Artificial intelligence used to evaluate job candidates must not become a tool that exacerbates discrimination.



By Alexandra Reeve Givens, Hilke Schellmann and Julia Stoyanovich

Ms. Givens is the chief executive of the Center for Democracy & Technology. Ms. Schellman and Dr. Stoyanovich are professors at New York University focusing on artificial intelligence.

https://www.nytimes.com/2021/03/17/opinion/ai-employment-bias-nyc.html

The measure must require companies to **publicly disclose what they find when they audit their tech for bias**. Despite pressure to limit its scope, the City Council must ensure that the bill would address discrimination in all forms — on the basis of not only race or gender but also disability, sexual orientation and other protected characteristics.

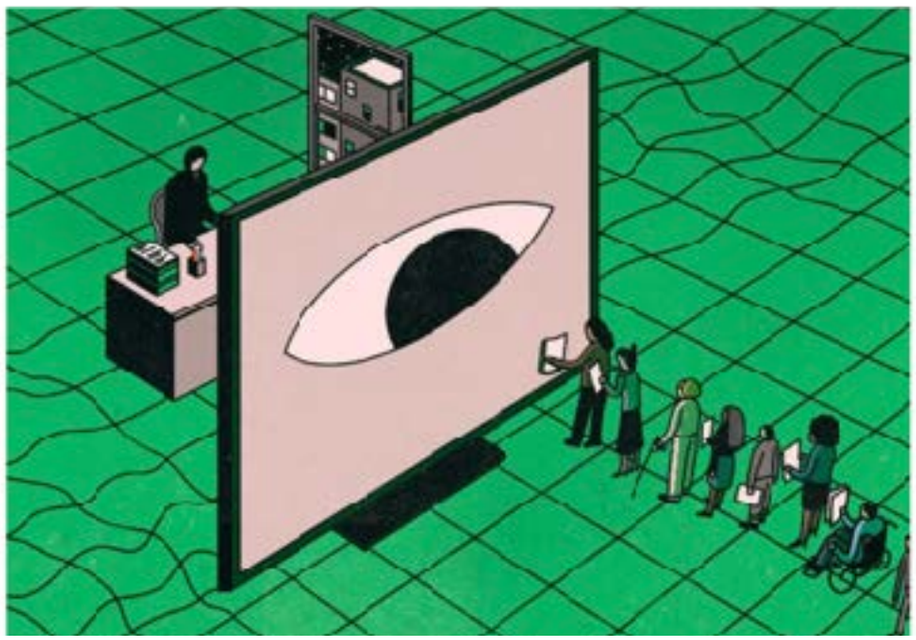
These audits should consider the circumstances of **people who are multiply marginalized** — for example, Black women, who may be discriminated against because they are both Black and women. Bias audits conducted by companies typically don't do this.

Hiring ADS regulation

The New York Times | March 17, 2021

We Need Laws to Take On Racism and Sexism in Hiring Technology

Artificial intelligence used to evaluate job candidates must not become a tool that exacerbates discrimination.



By Alexandra Reeve Givens, Hilke Schellmann and Julia Stoyanovich

Ms. Givens is the chief executive of the Center for Democracy & Technology. Ms. Schellman and Dr. Stoyanovich are professors at New York University focusing on artificial intelligence.

https://www.nytimes.com/2021/03/17/opinion/ai-employment-bias-nyc.html

The bill should […] require validity testing, to **ensure that the tools actually measure what they claim to**, and it must make certain **that they measure characteristics that are relevant for the job**. Such testing would interrogate whether, for example, candidates' efforts to blow up a balloon in an online game really indicate their appetite for risk in the real world — and whether risk-taking is necessary for the job.

… [T]he City Council must require vendors to tell candidates how they will be screened by an automated tool **before** the screening, so candidates know what to expect. People who are blind, for example, may not suspect that their video interview could score poorly if they fail to make eye contact with the camera. If they know what is being tested, they can engage with the employer to seek a fairer test.

Nutritional labels for job seekers

THE WALL STREET JOURNAL.

September 22, 2021

Hiring and AI: Let Job Candidates Know Why They Were Rejected

Labels that explain a hiring process that uses AI could allow job seekers to opt out if they object to the employer's data practices.

PHOTO: ISTOCKPHOTO/GETTY IMAGES

By Julia Stoyanovich

Updated Sept. 22, 2021 11:00 am ET

Artificial-intelligence tools are seeing ever broader use in hiring. But this practice is also hotly criticized because we rarely understand how these tools select candidates, and whether the candidates they select are, in fact, better qualified than those who are rejected.

To help answer these crucial questions, **we should give job seekers more information about the hiring process and the decisions**. The solution I propose is a twist on something we see every day: **nutritional labels**. Specifically, job candidates would see simple, standardized labels that show the factors that go into the AI's decision.

https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

Nutritional labels for job seekers

THE WALL STREET JOURNAL.

September 22, 2021

Hiring and AI: Let Job Candidates Know Why They Were Rejected

Labels that explain a hiring process that uses AI could allow job seekers to opt out if they object to the employer's data practices.

PHOTO: ISTOCKPHOTO/GETTY IMAGES

By Julia Stoyanovich

Updated Sept. 22, 2021 11:00 am ET

ACCOUNTANT

Acme Partners

Qualifications:	BS in accounting, GPA >3.0, Knowledge of financial and accounting systems and applications
Personal data to be analyzed:	An AI program could be used to review and analyze the applicant's personal data online, including LinkedIn profile, social media accounts and credit score.
Additional assessment:	AI-assisted personality scoring

ALERT: Applicants for this position DO NOT have the option to selectively decline use of AI analysis for any of their personal data or to review and challenge the results of such analysis.

https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

r/ai

New York City Local Law 144 of 2021

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

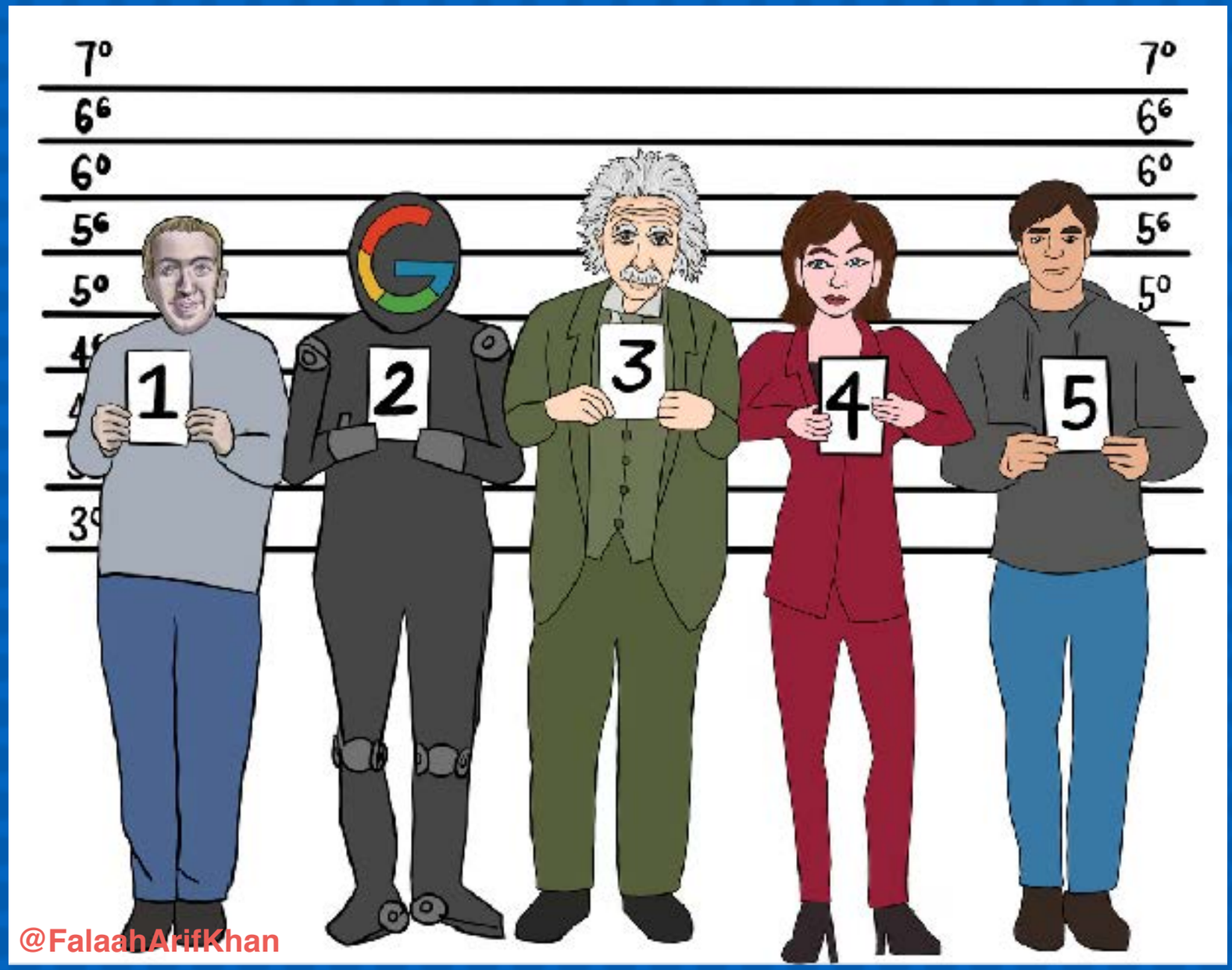
December 11, 2021

This bill would require that a **bias audit** be conducted on an automated employment decision tool prior to the use of said tool. The bill would also require that candidates or employees that reside in the city **be notified about the use of such tools** in the assessment or evaluation for hire or promotion, as well as, **be notified about the job qualifications and characteristics that will be used** by the automated employment decision tool. Violations of the provisions of the bill would be subject to a civil penalty.

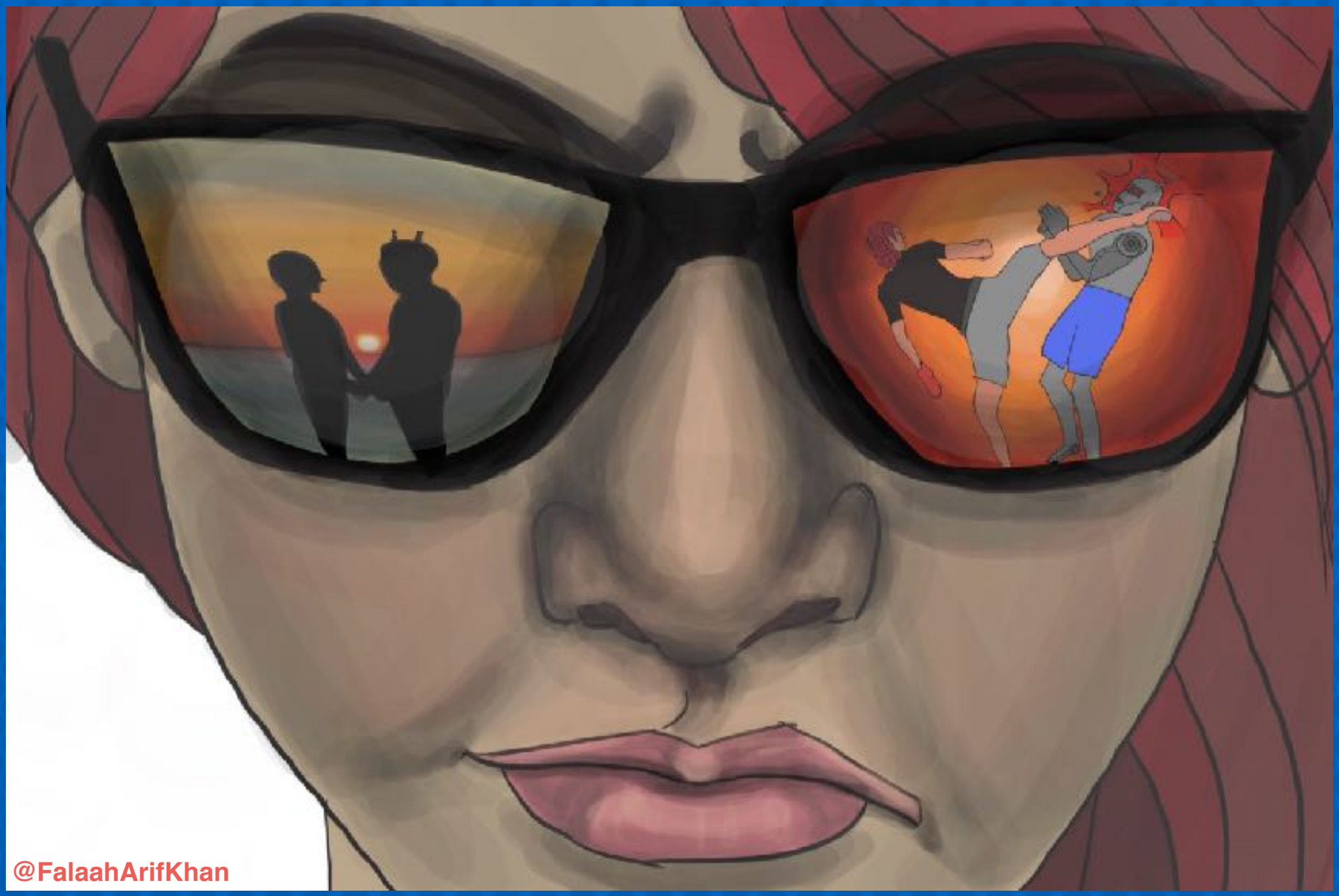
take-aways

r/ai

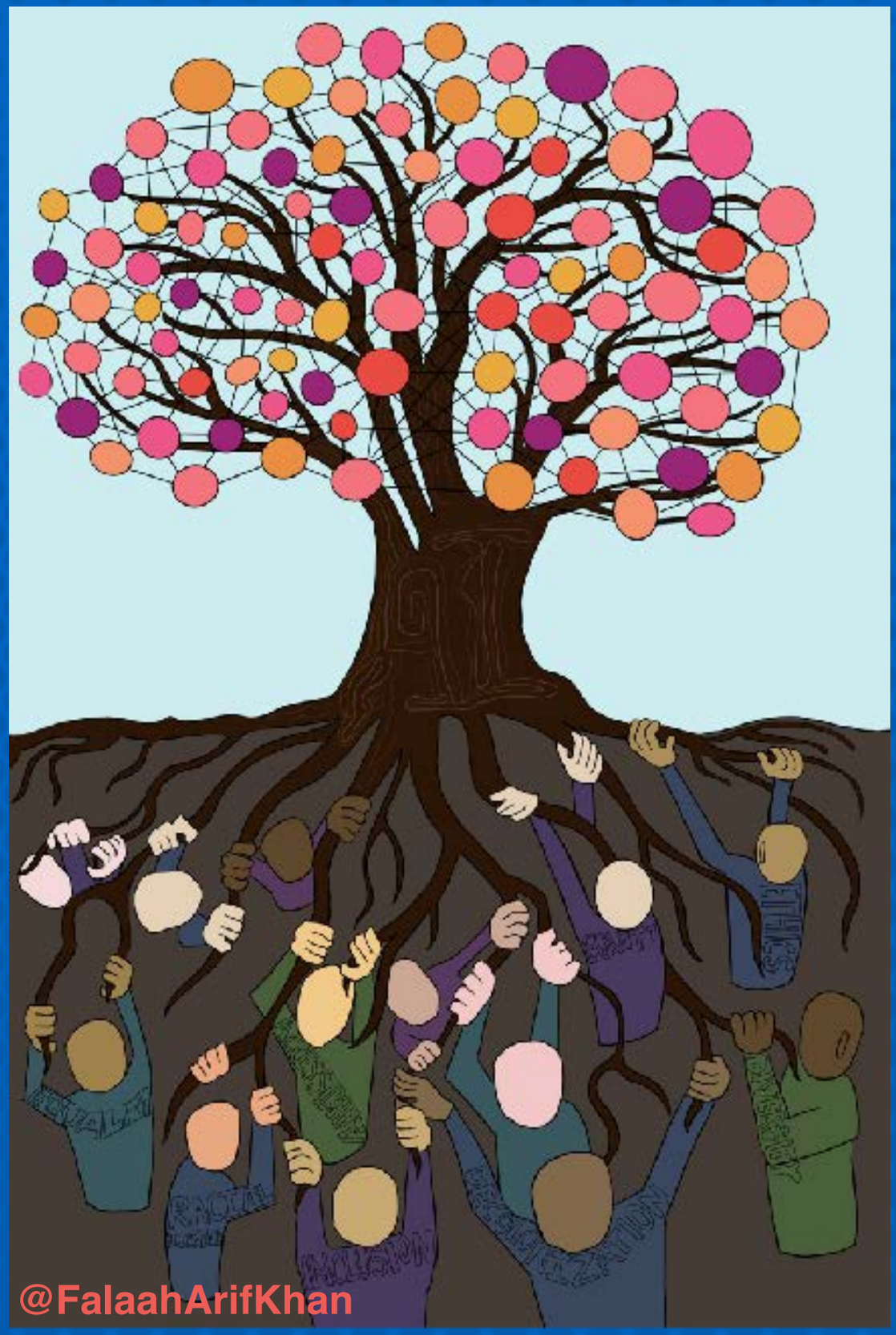
We all are responsible



Searching for balance

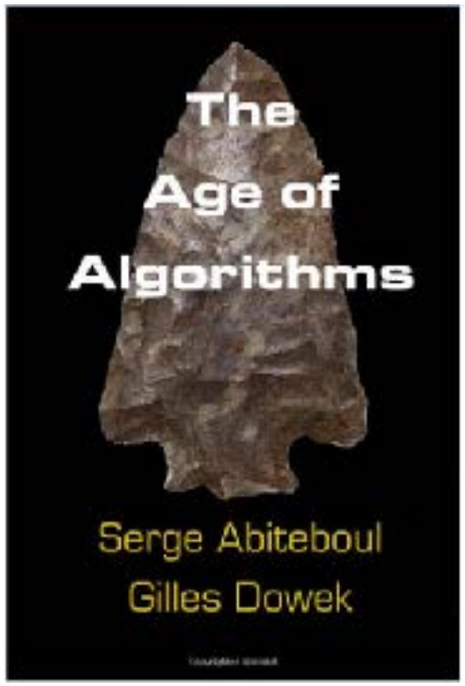


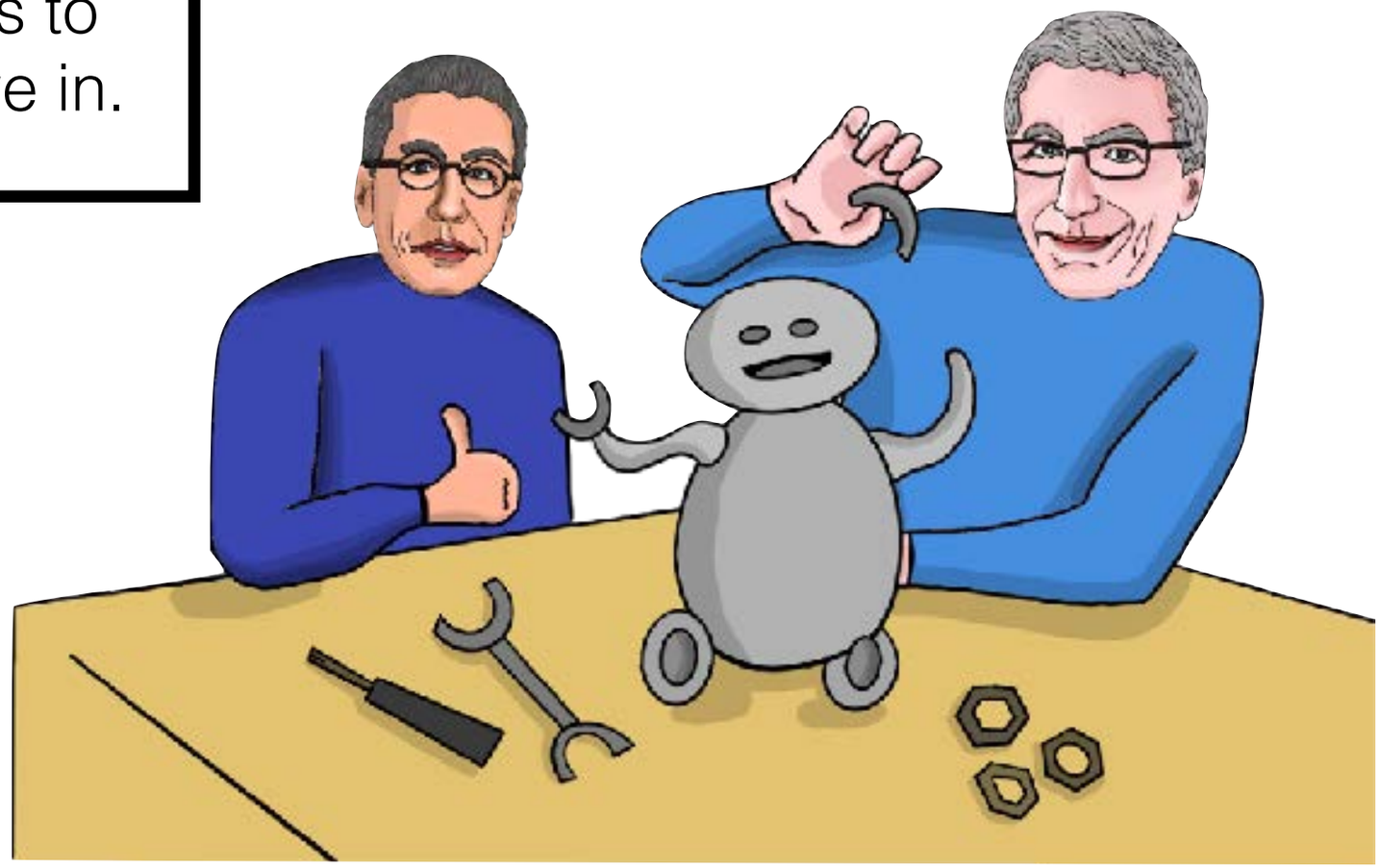
Tech rooted in people



AI is what *WE* make it!

Creations of the human spirit, **algorithms - and AI - are what we make them**. And they will be what we want them to be: it's up to us to choose the world we want to live in.





r/ai

Responsible Data Science

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai