Module: Fairness, Accountability, and Transparency of Algorithmic Systems

WEEKS 11-12: PRIVACY AND DATA PROTECTION (OVERVIEW)
TRAINRDM PROJECT
MAY 31ST, 2022

michael.bradford@ncirl.ie

Michael Bradford
michael.bradford@ncirl.ie

About Me

- Lecturer in the School of Computer Science at National College of Ireland
- Background in enterprise systems
- Areas of interest: Data Analytics, Machine Learning, Cloud Computing, Quantum Computing

Module aims and objectives

This module aims to provide learners with the knowledge and skills around the complex issues of data management and governance in an organisational context, including ethical and compliance issues that these present. Learners will explore the ethical, legal, and social implications of using data-driven technologies such as big data, analytics, internet of things, and machine learning. The students will learn how to establish processes and systems that consider best practices for data governance and adhere to ethical and regulatory requirements for data handling.

Minimum intended module learning outcomes

LO1 Demonstrate critical understanding of the governance and regulatory frameworks associated with the key data lifecycle stages for an effective management of data assets.

LO2 Demonstrate critical awareness and interpretation of the data privacy and data protection regulatory landscape in socio-technical environments.

LO3 Critically analyse and evaluate the main ethical, legal, and social implications of using data-driven technologies.

LO4 Investigate and appraise the interplay of fairness, accountability, and transparency in algorithmic decision-making systems and demonstrate awareness of technical solutions to enhance these concerns.

Agenda

Lecture Topic	Lecture Detail
Fairness, Accountability, and Transparency in Algorithmic Systems I	The meaning of fairness with respect to algorithmic systems. Techniques and models for fairness-aware data mining, information retrieval, recommendation, etc. Legal, social, and philosophical models of fairness. Specification of mathematical objectives with respect to fairness. Perceptions of algorithmic bias and unfairness. Interventions to mitigate biases in systems, or discourage biased behaviour from users.
Fairness, Accountability, and Transparency in Algorithmic Systems II	The meaning of accountability with respect to algorithmic systems. Processes and strategies for developing accountable systems. Methods and tools and standards for ensuring that algorithms comply with fairness policies (e.g., IEEE P7003 TM).
Fairness, Accountability, and Transparency in Algorithmic Systems III	The meaning of transparency with respect to algorithmic systems.. Explanations for algorithmic logic and outputs. Trade-offs between privacy and transparency. Tools and methodologies for conducting algorithm audits. Frameworks for conducting ethical and legal algorithm audits. Empirical results from algorithm audits.

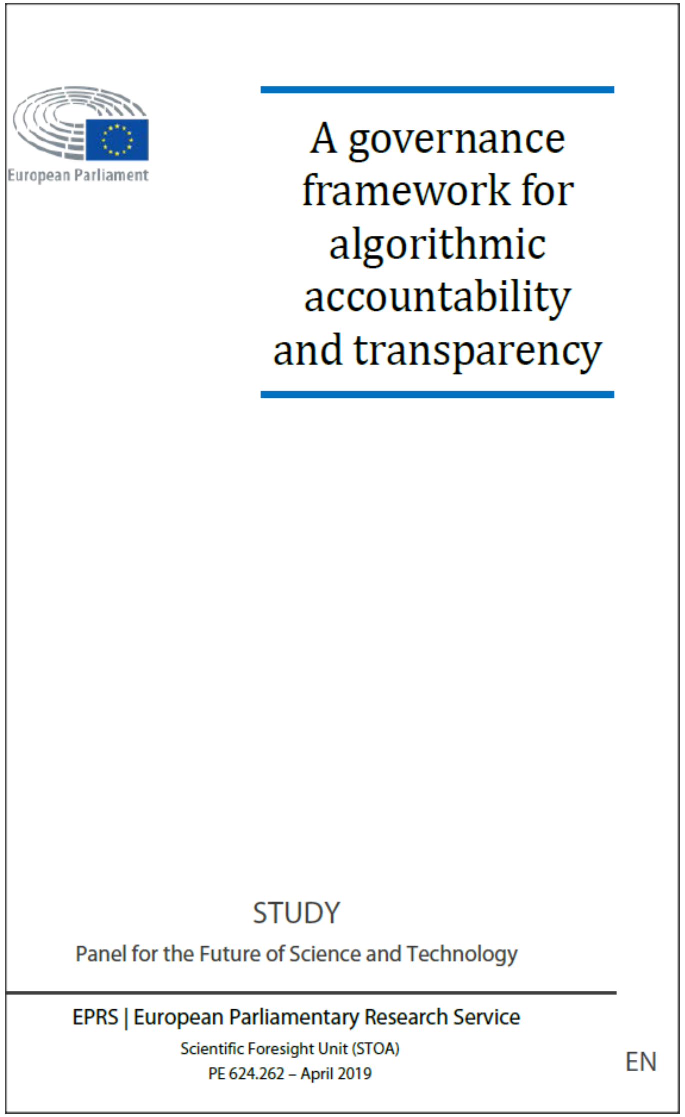
Fairness, Accountability, and Transparency (Part I)

Algorithmic fairness is of concern in Europe …

“Algorithmic systems are increasingly being used as part of decision-making processes in both the public and private sectors, with potentially significant consequences for individuals, organisations and societies as a whole.

… A significant factor in the adoption of algorithmic systems for decision-making is their capacity to process large amounts of varied data sets (i.e. big data), which can be paired with machine learning methods in order to infer statistical models directly from the data.

The same properties of scale, complexity and autonomous model inference however are linked to increasing concerns that many of these systems are opaque to the people affected by their use and lack clear explanations for the decisions they make.”



European Parliament

A governance framework for algorithmic accountability and transparency

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

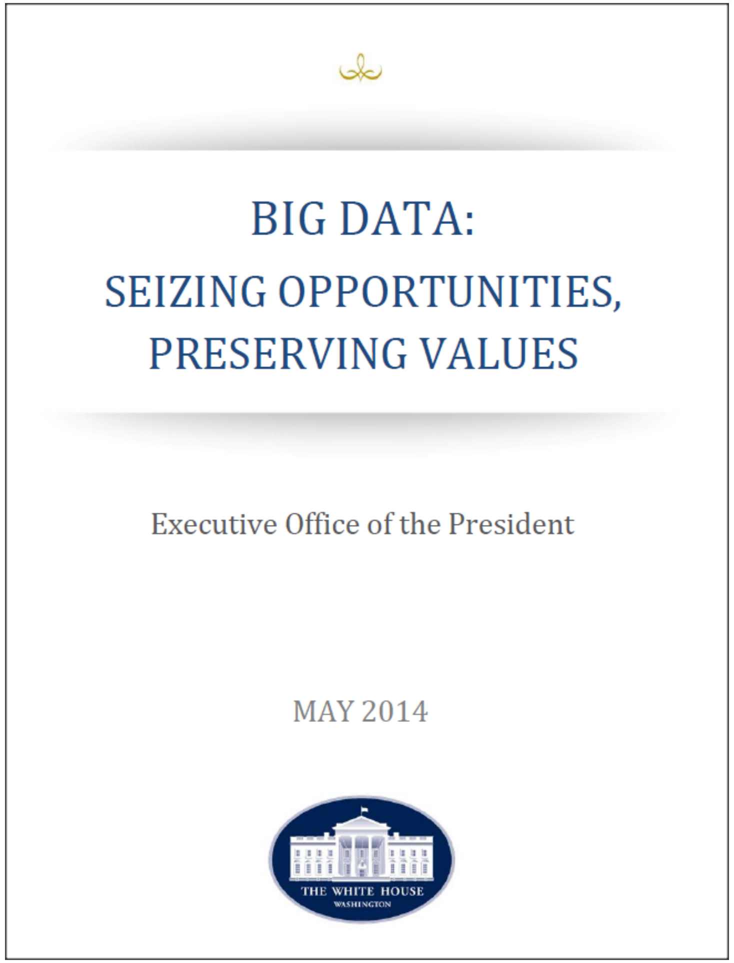
PE 624.262 – April 2019

EN

and the US …..

“In addition to creating tremendous social good, big data in the hands of government and the private sector can cause many kinds of harms. These harms range from tangible and material harms, such as financial loss, to less tangible harms, such as intrusion into private life and reputational damage.

An important conclusion of this study is that big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups. This discrimination can be the inadvertent outcome of the way big data technologies are structured and used. It can also be the result of intent to prey on vulnerable classes.”



BIG DATA:

SEIZING OPPORTUNITIES,
PRESERVING VALUES

Executive Office of the President

MAY 2014

big_data_privacy_report_may_1_2014.pdf (archives.gov)

… leading to common concerns

“This lack of transparency risks undermining meaningful scrutiny and accountability, which is a significant concern when these systems are applied as part of decision-making processes that can have a considerable impact on people's human rights (e.g., critical safety decisions in autonomous vehicles; allocation of health and social resources, etc.).”

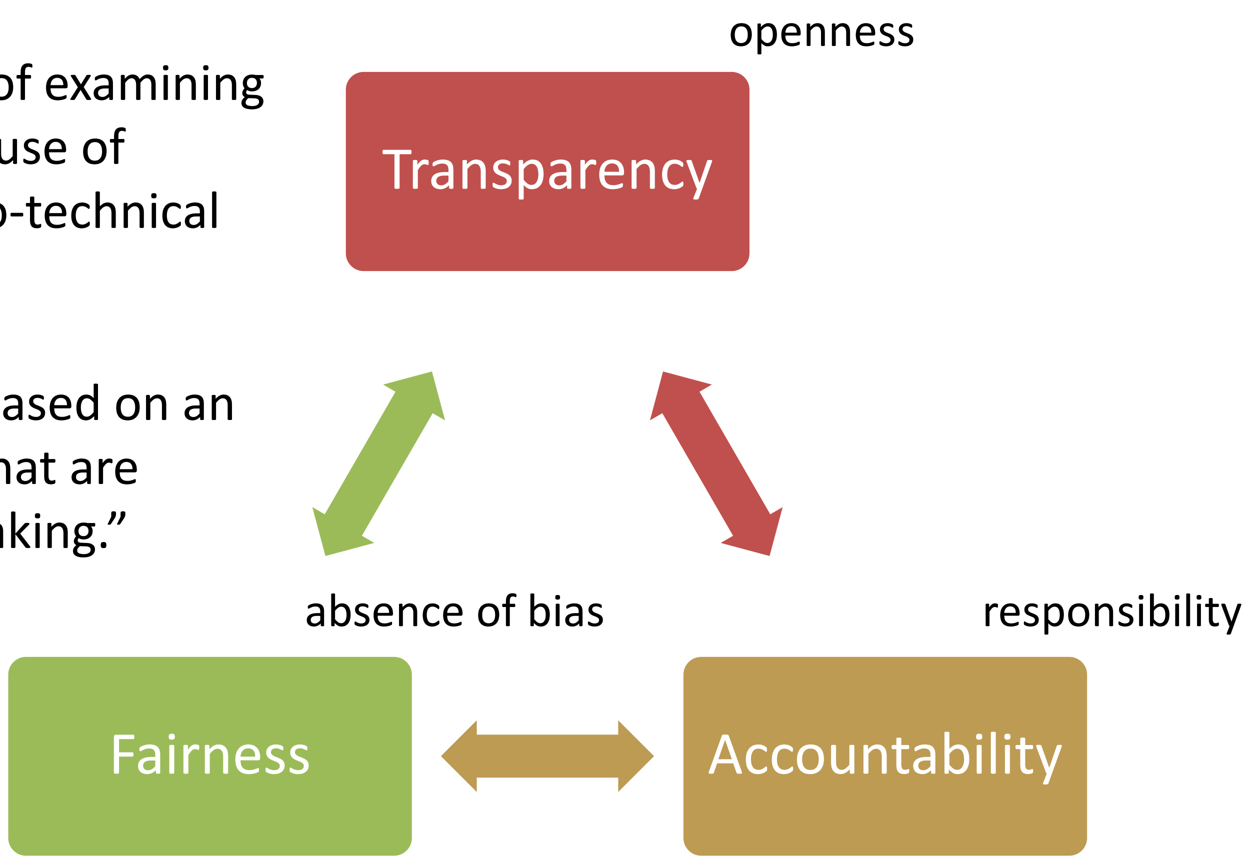
“Because of this lack of transparency and accountability, individuals have little recourse to understand or contest the information that has been gathered about them or what that data, after analysis, suggests. … the civil rights community is concerned that such algorithmic decisions raise the specter of “redlining” in the digital economy—the potential to discriminate against the most vulnerable classes of our society under the guise of neutral algorithms.”

Fairness, Accountability and Transparency of Algorithms

Fairness, Accountability and Transparency – way of examining critically one of the principal concerns about the use of algorithms in society (as part of the broader socio-technical system)

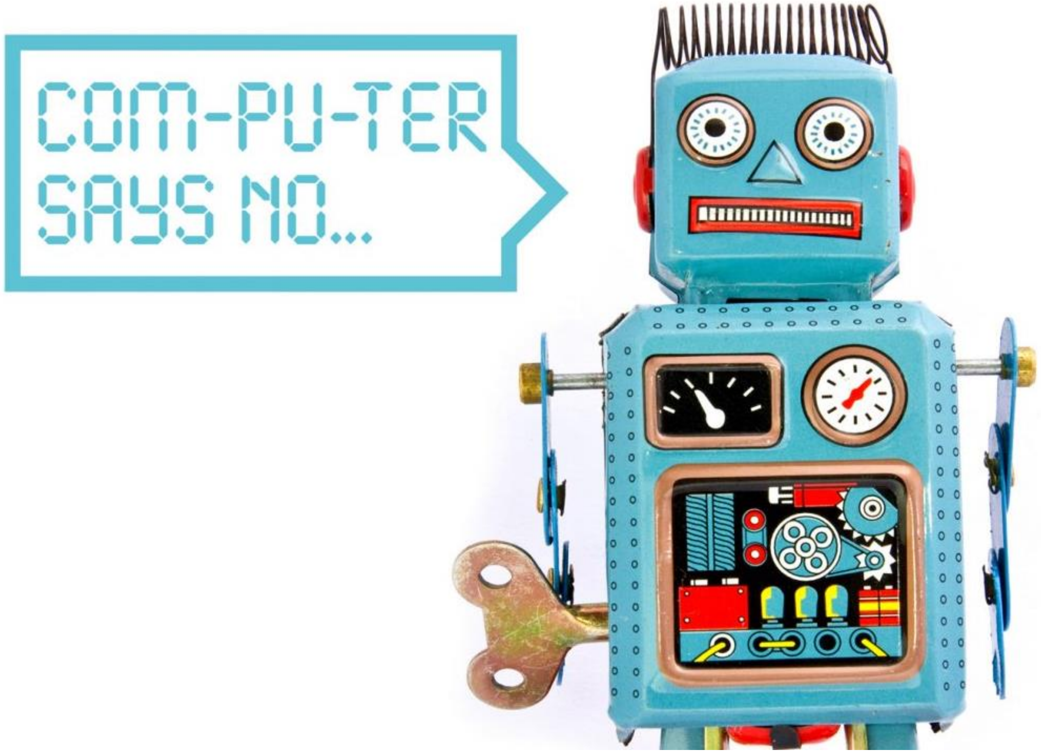
“In principle, **fairness** is the absence of any bias based on an individual’s inherent or acquired characteristics that are irrelevant in the particular context of decision-making.”

- To assess fairness, we need transparency
- To enforce fairness, we need accountability



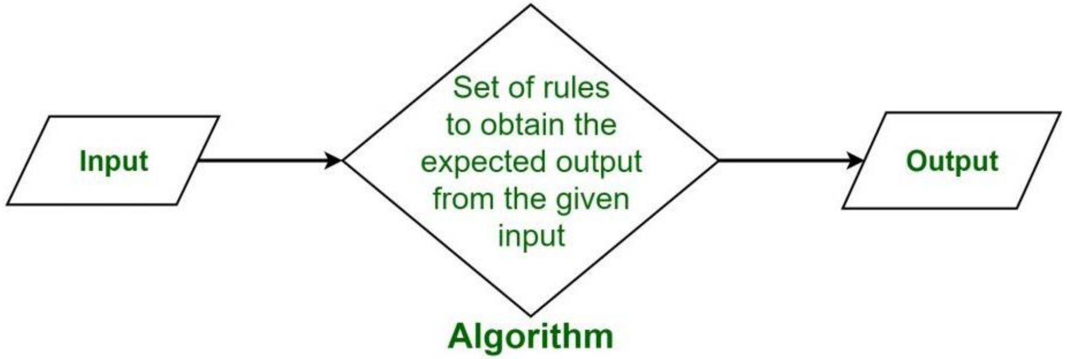
An Algorithmic System

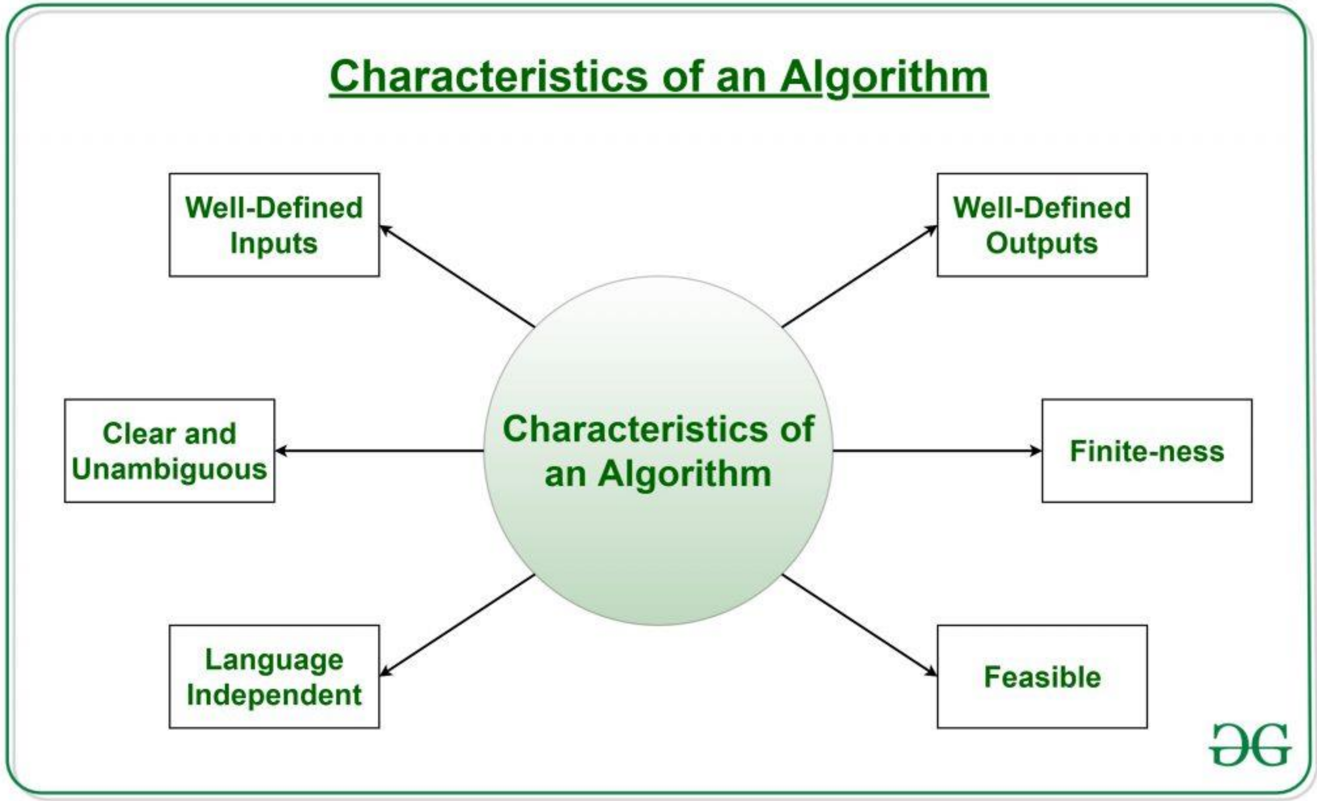
- An **algorithmic system** is a **system** comprised of one or more algorithms used in a software to collect and analyze data as well as draw conclusions as part of a process designed to solve a pre-**defined** problem
- Algorithmic decision-making systems (AD-M sys) are ubiquitous across a wide variety of services. They rely on:
  - on complex learning methods
  - vast amounts of data
- There is a growing concern that these automated decisions can lead to a lack of fairness



An Algorithmic System

- **"Algorithm"** refers to a set of precise instructions or rules regarding actions to be taken in solving a predefined problem


Input
Set of rules to obtain the expected output from the given input
Output
Algorithm



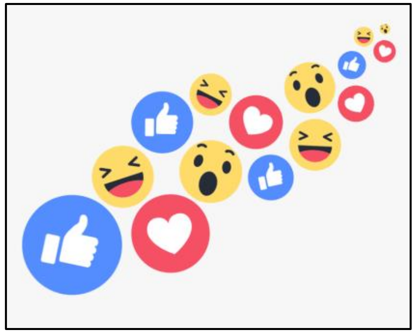
- **An algorithmic system** is **a system** comprised **of one or more algorithms** used in a software to collect and analyze data as well as draw conclusions **as part of a process** designed to solve a pre-defined problem

The Use of Algorithms

Algorithms can be used for:

- Mapping your online world

To predict what customers might want to buy/watch



To conduct sentiment analysis of posts

To enable smart assistants

- Analyzing what is in your personal filter bubble

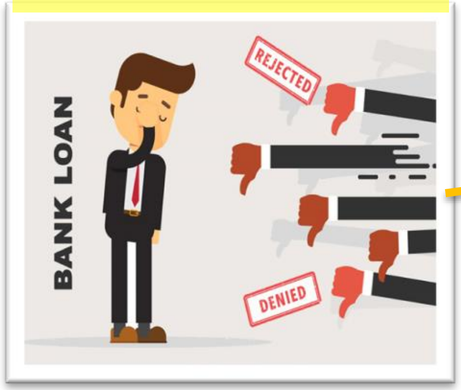
Not everyone gets the same results online

AMAZON
HUFFINGTON POST
NETFLIX
YAHOO NEWS
YOU
GOOGLE
FLIPBOARD
TARGETED ADS
PERSONAL DATA

The Use of Algorithms

What happens when the output of decision-making algorithms have significant societal impact?

Loans



Issues have arisen regarding the fairness of these algorithmic decisions

Hiring

University Admissions

Bails

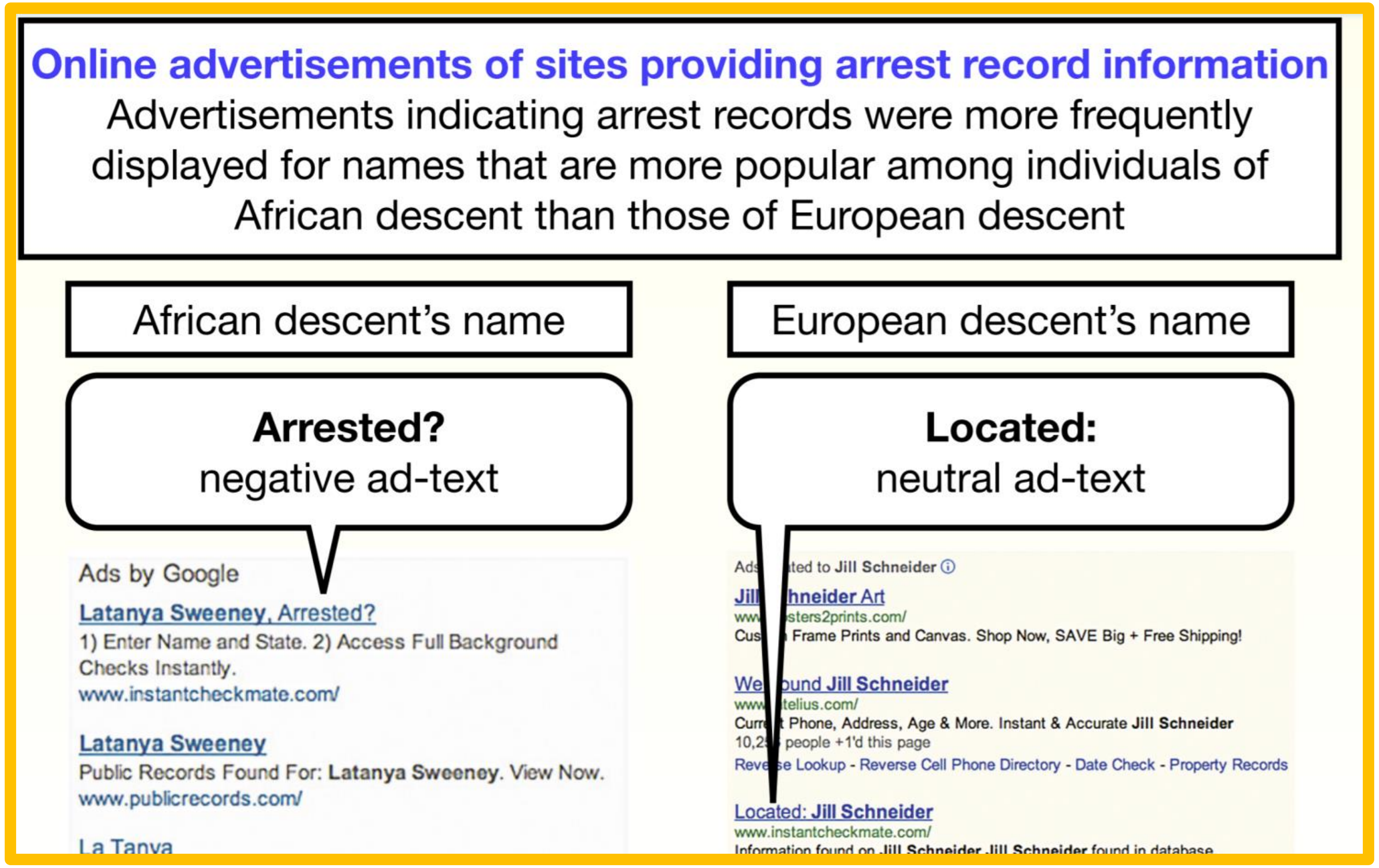
Examples Of Algorithmic Biases

Bias in online ads

Latanya Sweeney, Harvard researcher and former chief technology officer at the Federal Trade Commission (FTC), found that

online search queries for African-American names were more likely to return ads to that person from a service that renders arrest records, as compared to the ad results for white names.

the same differential treatment occurred in the micro-targeting of higher-interest credit cards and other financial products when the computer inferred that the subjects were African-Americans, despite having similar backgrounds to whites.



Online advertisements of sites providing arrest record information

Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent

African descent's name

Arrested?
negative ad-text

Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya Sweeney**. View Now.
www.publicrecords.com/

La Tanya

European descent's name

Located:
neutral ad-text

Ads ted to **Jill Schneider**

Jill hneider Art
www sters2prints.com/
Cus Frame Prints and Canvas. Shop Now, SAVE Big + Free Shipping!

We ound **Jill Schneider**
www telius.com/
Curr t Phone, Address, Age & More. Instant & Accurate **Jill Schneider**
10,2 people +1'd this page
Reve se Lookup - Reverse Cell Phone Directory - Date Check - Property Records

Located: **Jill Schneider**
www.instantcheckmate.com/
Information found on **Jill Schneider**. **Jill Schneider** found in database

Sweeney, Latanya. "Discrimination in online ad delivery." Rochester, NY: Social Science Research Network, January 28, 2013. Available at https://papers.ssrn.com/abstract=2208240 (last accessed April 12, 2019).

Examples Of Algorithmic Biases

Bias in facial recognition technology

the algorithms powering three commercially available facial recognition software systems were failing to recognize darker-skinned complexions

most facial recognition training data sets are estimated to be more than 75% male and more than 80% white.

When the person in the photo was a white man, the software was accurate 99% of the time at identifying the person as male.

Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." MIT News, February 11, 2018. Available at http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed April 19, 2019).

Bias in criminal justice algorithms

COMPAS algorithm, which is used by judges to predict whether defendants should be detained or released on bail pending trial, was found to be biased against African-Americans

The algorithm assigns a risk score to a defendant's likelihood to commit a future offense, relying on the voluminous data available on arrest records, defendant demographics, and other variables.

Compared to whites who were equally likely to re-offend, African-Americans were more likely to be assigned a higher-risk score, resulting in longer periods of detention while awaiting trial

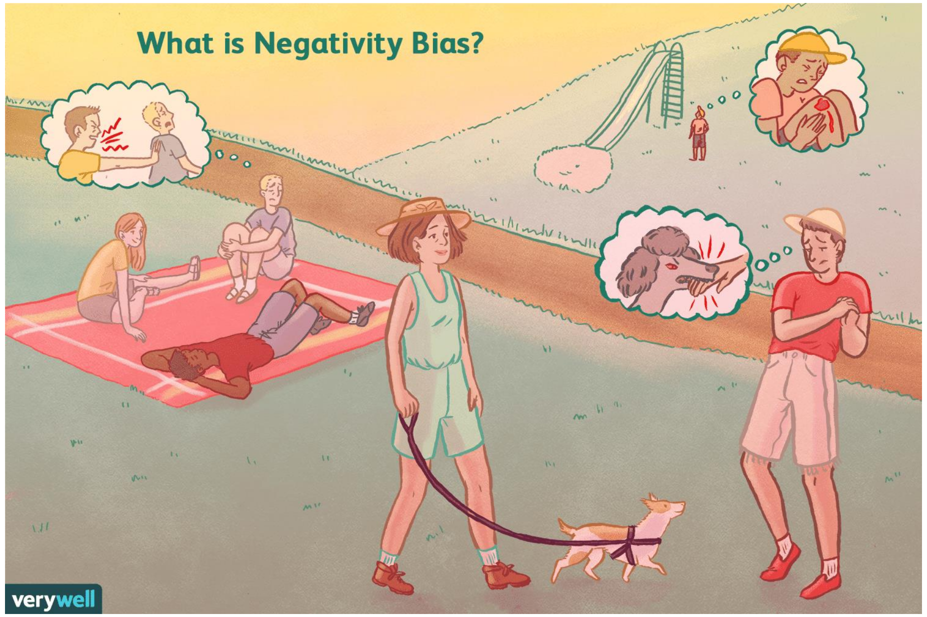
Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness." ArXiv:1701.08230 [Cs, Stat], January 27, 2017. https://doi.org/10.1145/3097983.309809

Historical Human Biases

Historical human biases are shaped by pervasive and often deeply embedded prejudices against certain groups, which can lead to their reproduction and amplification in computer models.

- As humans, we tend to:
  - a) Remember traumatic experiences better than positive ones
  - b) Recall insults better than praise.
  - c) React more strongly to negative stimuli.
  - d) Think about negative things more frequently than positive ones.
  - e) Respond more strongly to negative events than to equally positive ones.

- these realities will be reflected in the training data
- If historical biases are factored into the model, it will make the same kinds of wrong judgments that people do.

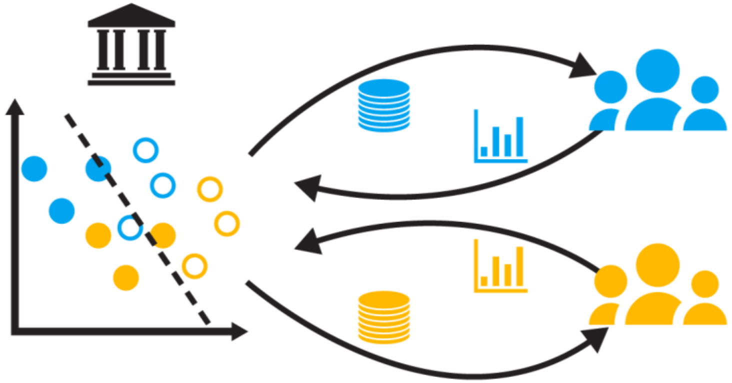


Results of Historical Human Biases

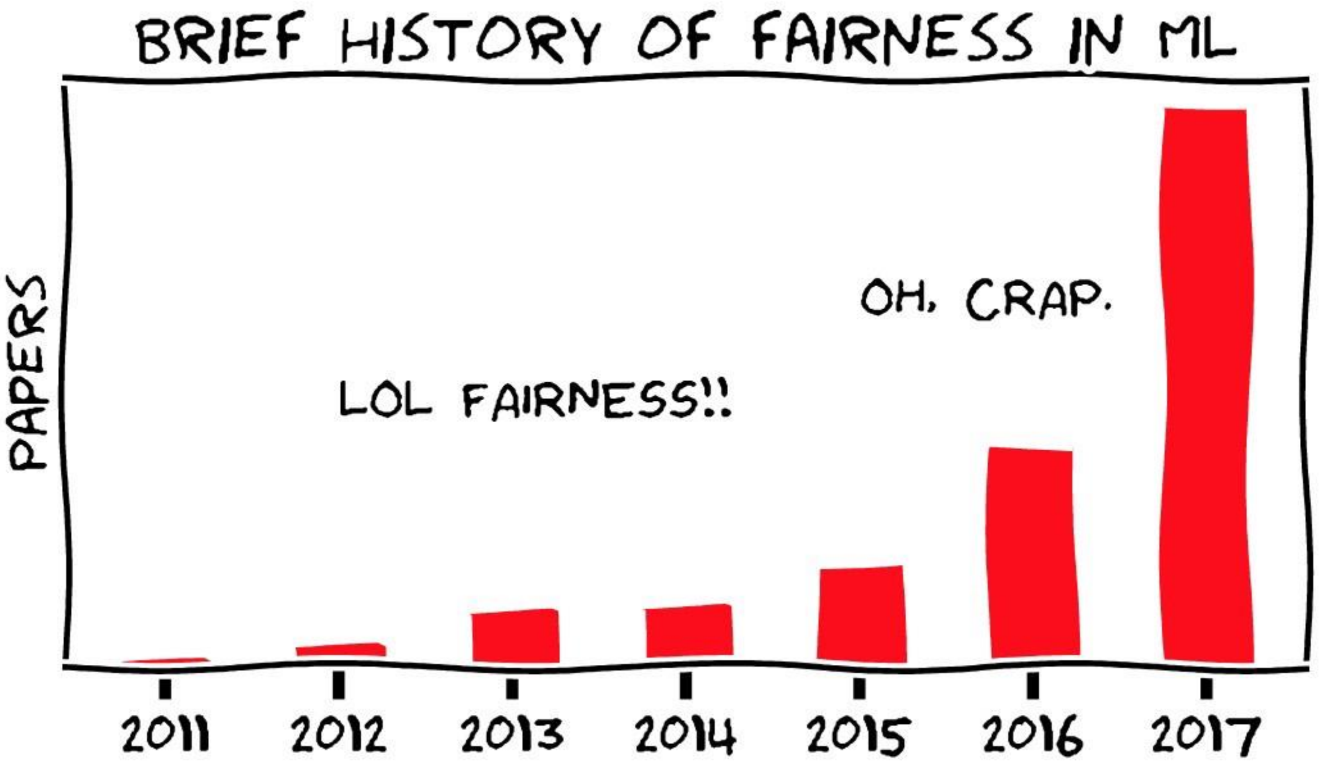
Further, human biases can be reinforced and perpetuated without the user's knowledge.

For example, African-Americans who are primarily the target for high-interest credit card options might find themselves clicking on this type of ad without realizing that they will continue to receive such predatory online suggestions. In this and other cases, the algorithm may never accumulate counter-factual ad suggestions (e.g., lower-interest credit options) that the consumer could be eligible for and prefer.

Thus, it is important for algorithm designers and operators to watch for such potential negative feedback loops that **cause an algorithm to become increasingly biased over time**.



Fairness in Machine Learning Research



The number of publications on fairness from 2011 to 2017

ICML | 2022

Thirty-ninth International Conference on Machine Learning

Fairness, Accountability, and Transparency in Machine Learning

ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

FairWare 2018

International Workshop on Software Fairness

May 29, 2018
Gothenburg, Sweden
Collocated with ICSE 2018

Definition of Fairness

How People Perceive The Fairness?

In principle, fairness is the absence of any bias based on an individual's inherent or acquired characteristics that are irrelevant in the particular context of decision-making

#1 **Treating similar individuals similarly**: This is determined by a similarity distance metric (applied to certain attributes) which represents a notion of ground truth in regard to the decision context.

->Individuals with similar repayment rates should receive similar amounts of money

#2 **Never favor a worse individual over a better one**: This definition promotes meritocracy with respect to the candidate's inherent quality

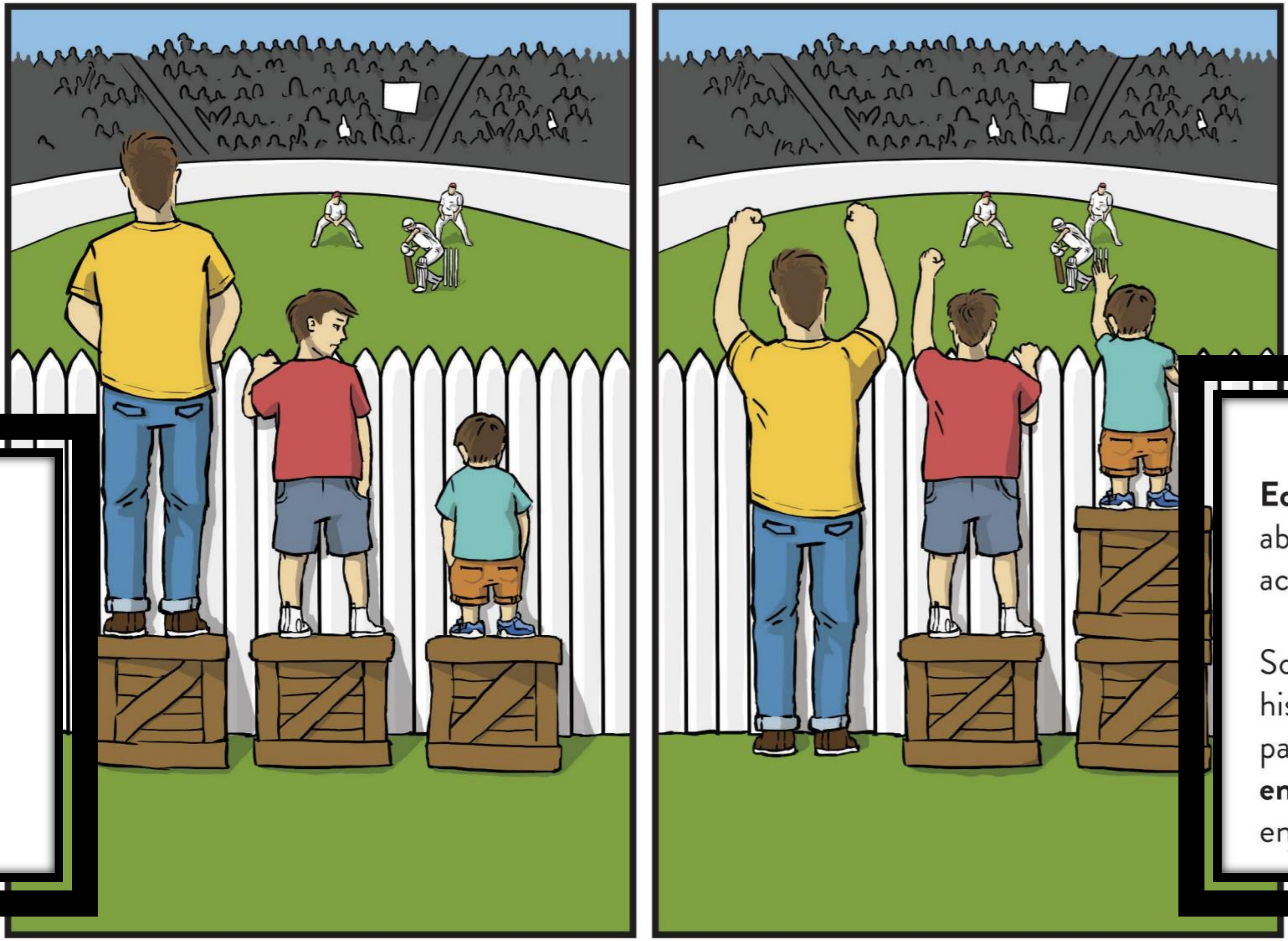
-> An individual with a higher repayment rate should obtain at least as much money as her peer

#3 **Calibrated fairness**: selects individuals in proportion to their merit

-> two individuals with repayment rates r1 and r2, respectively, should obtain r1/(r1 + r2) and r2/(r1 + r2) amount of money, respectively

Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C. and Liu, Y., 2019, January. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 99-106).

What Is Fairness?



Equality = SAMENESS

Equality is about SAMENESS, it promotes fairness and justice by giving everyone the same thing.

BUT it can **only work IF everone starts from the SAME place**, in this example equality only works if everyone is the same height.

Equity = FAIRNESS

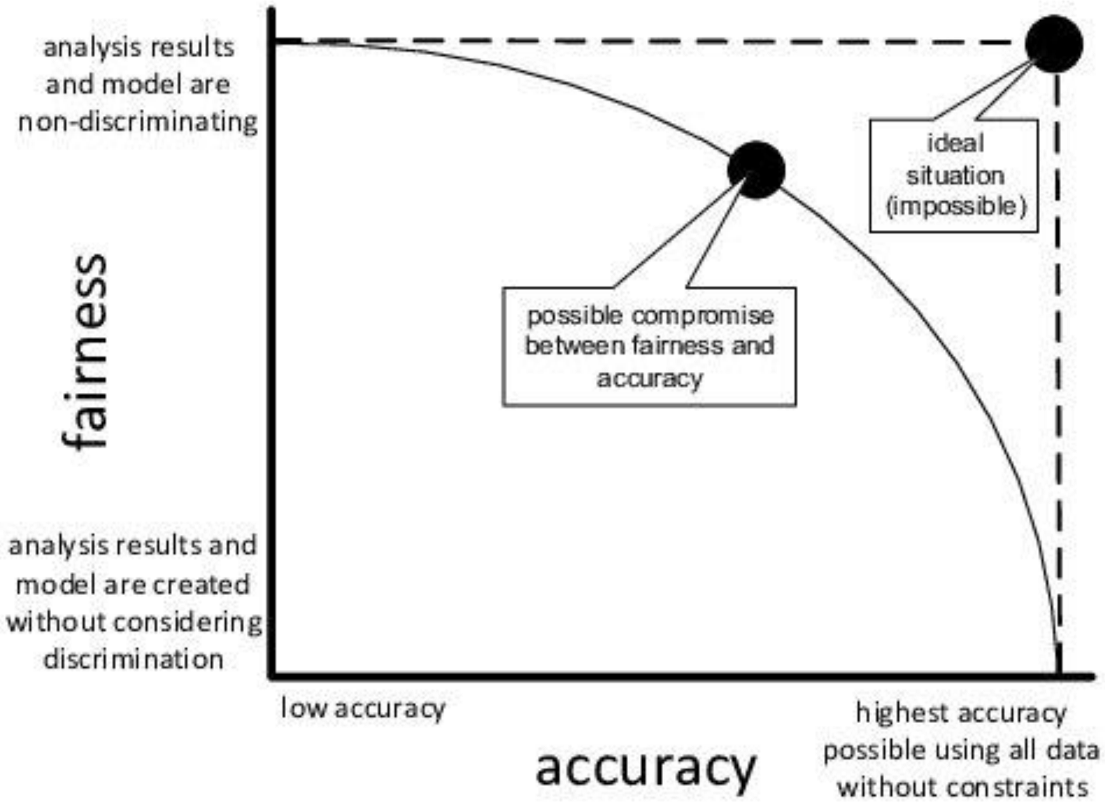
Equity is about FAIRNESS, it's about making sure people get access to the same opportunities.

Sometimes our differences and/or history can make barriers to participation, so we must **FIRST ensure EQUITY** before we can enjoy equality.

The Meaning Of Fairness With Respect To Algorithmic Systems

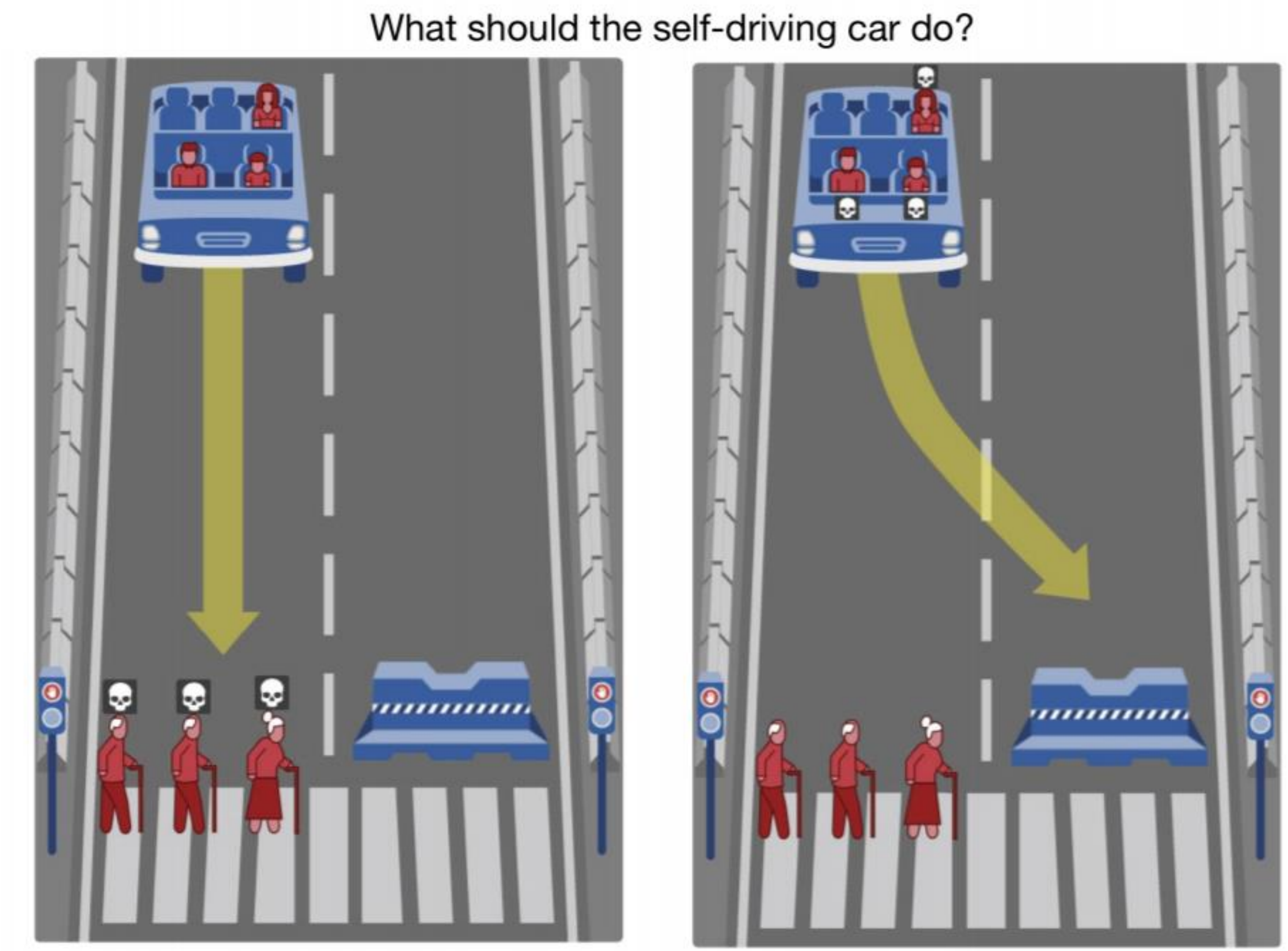
Studying *fairness in ML decision-making algorithms* means studying the algorithms **not only from a perspective of accuracy, but also from a perspective of fairness**

- The difficulty revolves around **defining what fairness** means
- In many cases these definitions have trade-offs with accuracy (i.e, achieving them means necessarily paying a price in terms of the model's accuracy)


analysis results and model are non-discriminating
fairness
analysis results and model are created without considering discrimination
ideal situation (impossible)
possible compromise between fairness and accuracy
low accuracy
accuracy
highest accuracy possible using all data without constraints

Van Der Aalst, W.M., 2016. Green data science: using big data in an" environmentally friendly" manner. In *18th international conference on enterprise information systems (ICEIS 2016)* (pp. 9-21). SCITEPRESS-Science and Technology Publications, Lda..

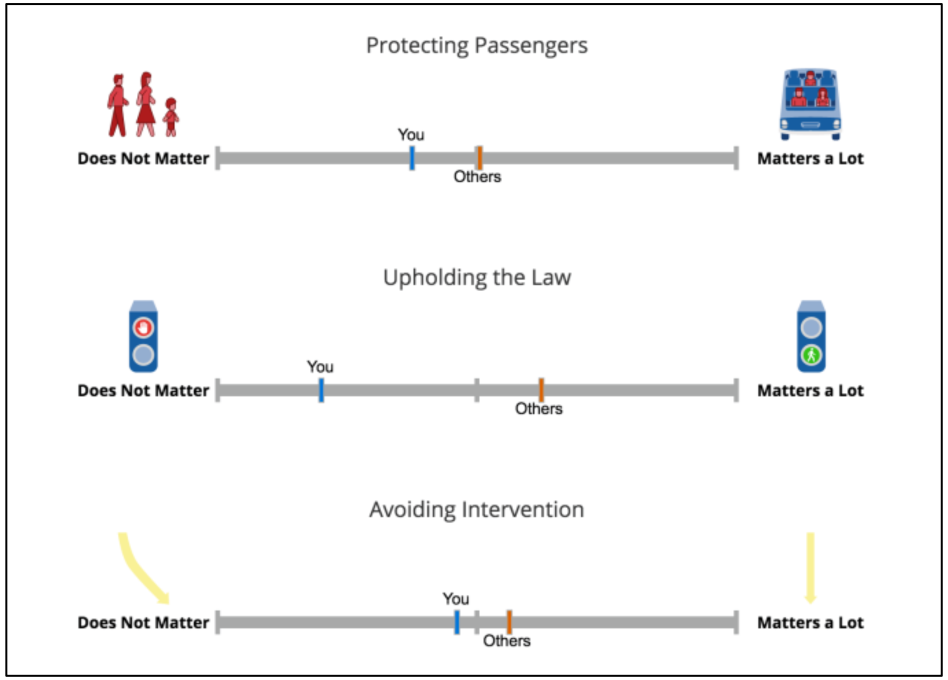
The Moral Machine Experiment



What should the self-driving car do?

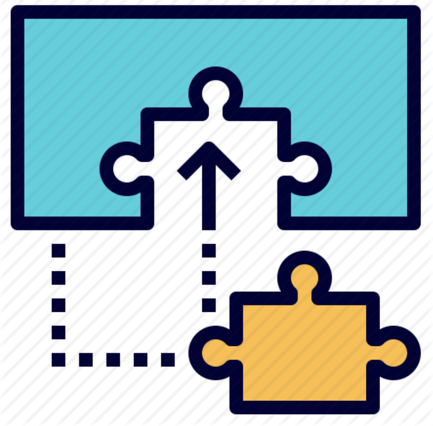
An online experimental platform designed to gather a human perspective on moral decisions faced by autonomous vehicles

Allowing machines to choose whether to kill humans would be devastating for world peace and security



https://www.moralmachine.net/

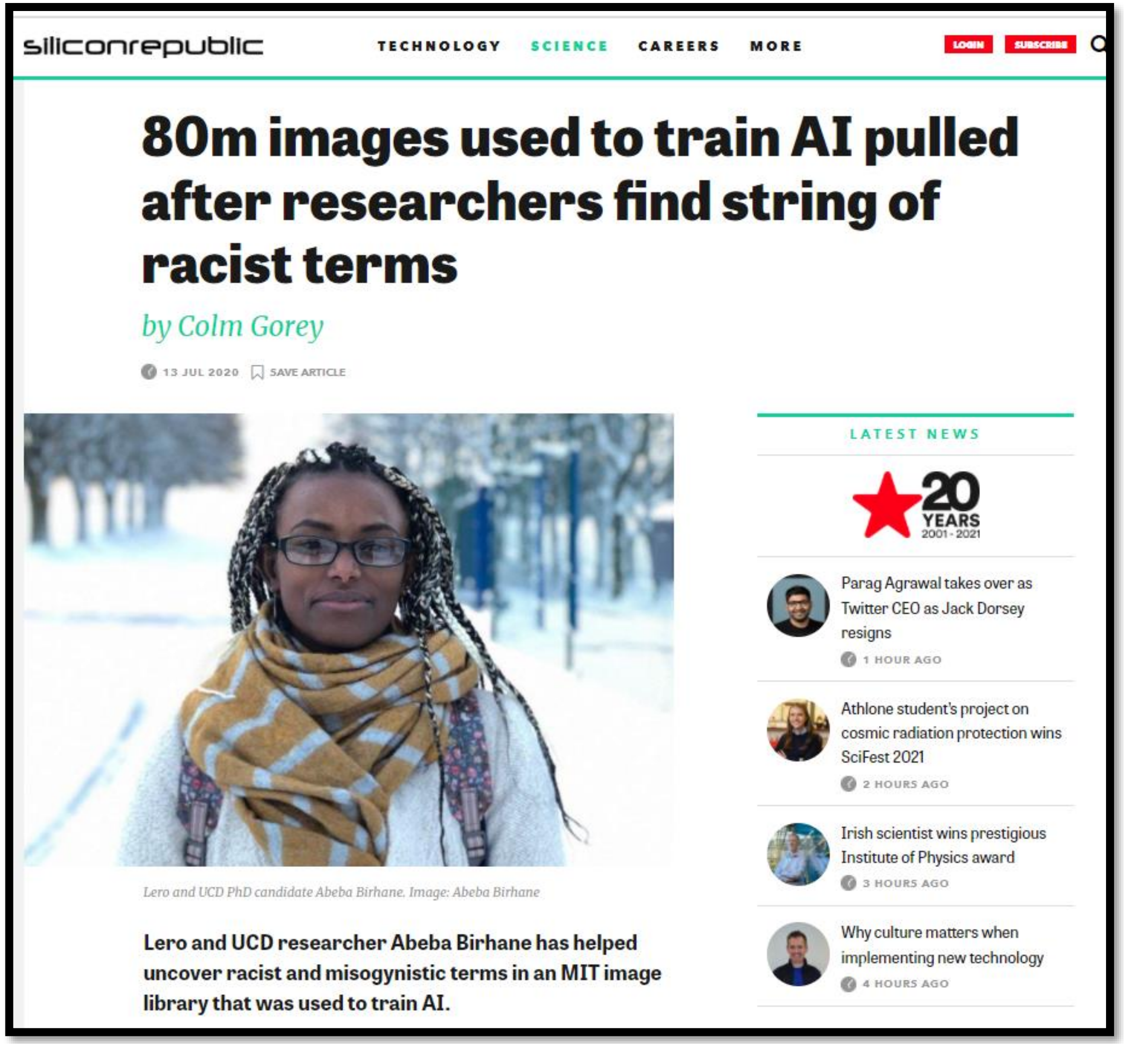
Potential Causes Of Unfairness



1. Biases **already included in the datasets**, which are based on:
   - •biased device measurements
   - •historically biased human decisions - unfair labelling by annotators
   - •erroneous reports or other reasons
2. Biases caused by **missing data**
3. Sample selection bias - datasets are not representative of the target population
4. **Limited features** - features may be less informative or reliably collected for minority group
5. Size disparity – **unbalanced dataset**
6. Biases caused by **“proxy" attributes** (non-sensitive attributes that can be exploited to derive sensitive attribute)

Pessach, D. and Shmueli, E., 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Example of problems in datasets



80m images used to train AI pulled after researchers find string of racist terms

by Colm Gorey

13 JUL 2020 SAVE ARTICLE

Lero and UCD PhD candidate Abeba Birhane. Image: Abeba Birhane

Lero and UCD researcher Abeba Birhane has helped uncover racist and misogynistic terms in an MIT image library that was used to train AI.

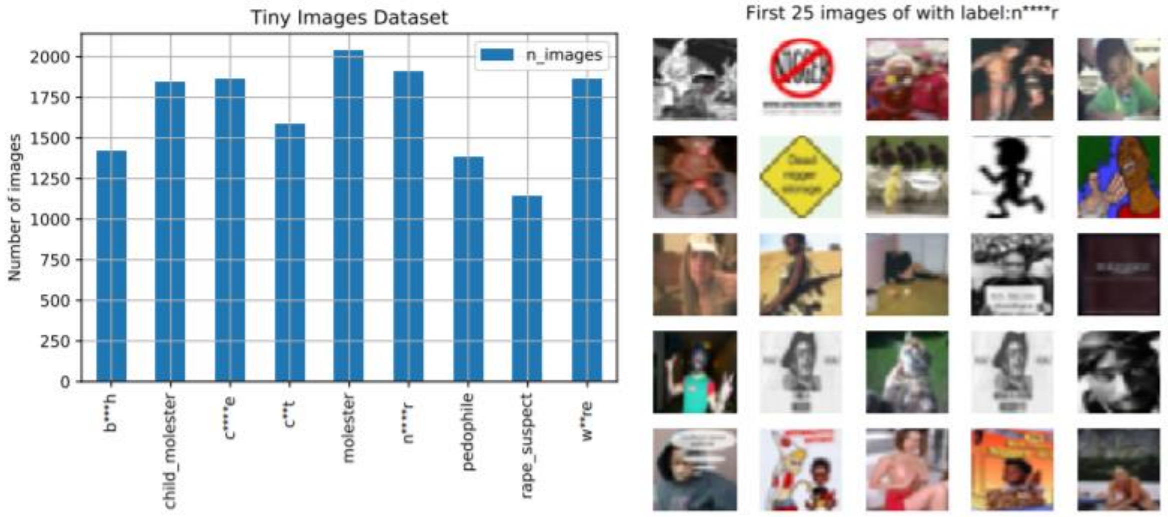
LATEST NEWS

Parag Agrawal takes over as Twitter CEO as Jack Dorsey resigns
1 HOUR AGO

Athlone student's project on cosmic radiation protection wins SciFest 2021
2 HOURS AGO

Irish scientist wins prestigious Institute of Physics award
3 HOURS AGO

Why culture matters when implementing new technology
4 HOURS AGO



(a) Class-wise counts of the offensive classes

(b) Samples from the class labelled n****r

Figure 1: Results from the *80 Million Tiny Images* dataset exemplifying the toxicities of it's label space

Birhane, A. and Prabhu, V.U., 2021, January. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1536-1546). IEEE.

Detecting And Mitigating Bias (1/4)

Bias detection should begin with **careful handling of the sensitive information of users**, including data that identify a person's membership in a federally protected group (e.g., race, gender).

In some cases, operators of algorithms may also worry about a person's membership in some other group if they are also susceptible to unfair outcomes.

An examples of this could be college admission officers worrying about the algorithm's exclusion of applicants from lower-income or rural areas; these are individuals who may be not federally protected but do have susceptibility to certain harms (e.g., financial hardships).

Recent research has proposed to **use encryption over sensitive attributes** to improve fairness while still allowing the model to be trained, checked, or have its outputs verified and held to account

Detecting And Mitigating Bias (2/4)

Computer programmers normally examine the set of outputs that the algorithm produces to **check for anomalous results**.

- Comparing outcomes for different groups can be a useful first step.

This could even be done through simulations

- companies consider the simulation of predictions (both true and false) before applying them to real-life scenarios

For example, if a job-matching algorithm's average score for male applicants is higher than that for women, further investigation and simulations could be warranted.

Detecting And Mitigating Bias (3/4)

- **Fairness-Aware Data Collection and Curation**
  - improved attention to identifying blind spots, biases and limitations in machine learning team and other actors (such as those responsible for outcome decisions /labelling)

- More proactive and holistic auditing processes

- Addressing issues detected during machine learning

Holstein, K. *et al.* (2019) ‘Improving fairness in machine learning systems: What do industry practitioners need?’, *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–16.

Detecting And Mitigating Bias (4/4)

- Ethical Frameworks
    - **Ethics Guidelines for Trustworthy AI**: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, nondiscrimination and fairness, (6) environmental and societal well-being, and (7) accountability
    - it is unethical to “unfairly discriminate”

- Algorithmic Impact Assessments (AIAs)

Legal, Social, And Philosophical Models Of Fairness

Quantitative restrictions by regulations or laws against discrimination:

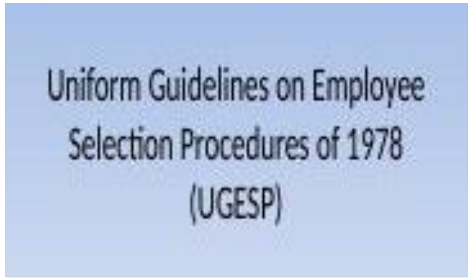
Racial Equality Directive of E.U. shall be taken to occur where one person is treated less favorably than another is in a comparable situation on grounds of racial or ethnic origin

Uniform Guidelines on Employee Selection Procedure (US) a selection rate for any race, sex, or ethnic group which is less than four-fifths (or 80%) of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact

Anti-Discrimination Act (Australia, Queensland) a person treats, or proposes to treat, a person with an attribute less favorably than another person without the attribute

FRA

The **Racial Equality Directive:** application and challenges



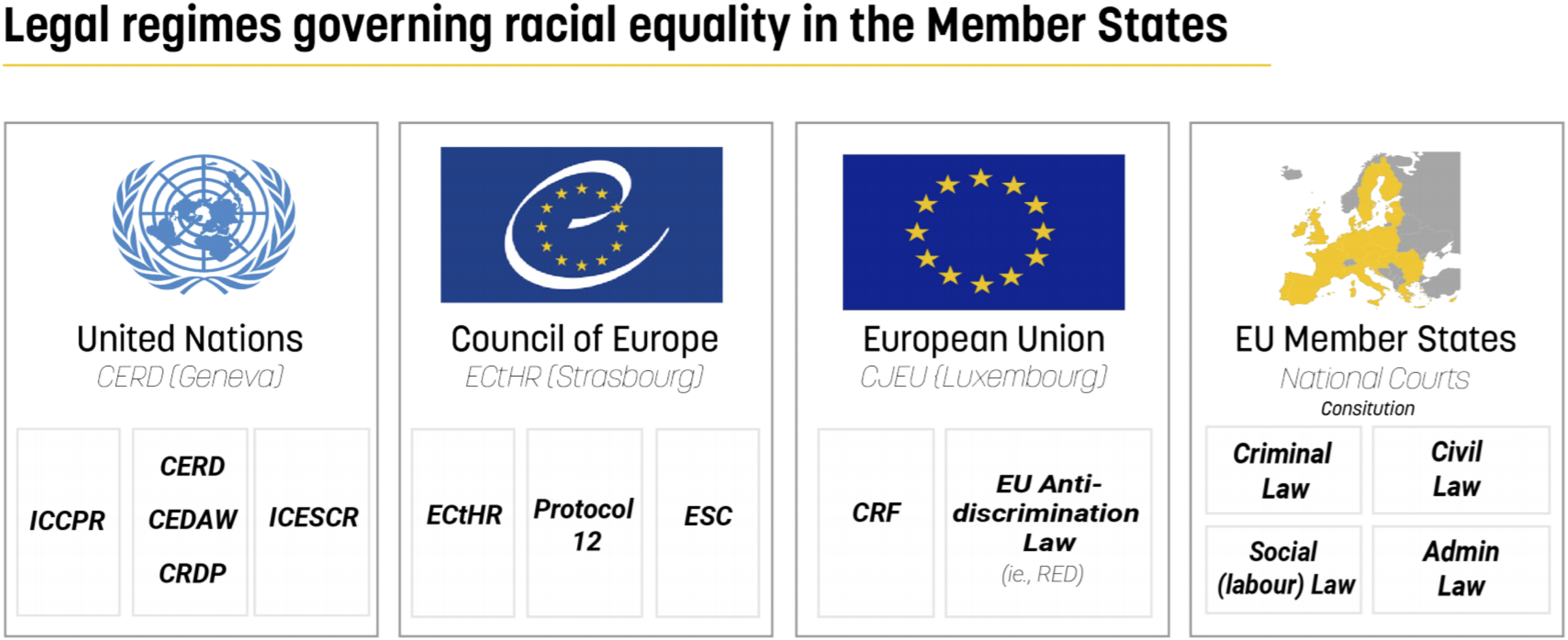
Uniform Guidelines on Employee Selection Procedures of 1978 (UGESP)

Racial Discrimination Act 1975

Act No. 52 of 1975 as amended

D. Pedreschi, S. Ruggieri, and F. Turini. Measuring discrimination in socially-sensitive decision records. In Proc. of the SIAM Int’l Conf. on Data Mining, pages 581–592, 2009. doi: https: //doi.org/10.1137/1.9781611972795.50

Legal, Social, And Philosophical Models Of Fairness



Legal regimes governing racial equality in the Member States

United Nations
CERD (Geneva)

- ICCPR
- CERD
CEDAW
CRDP
- ICESCR

Council of Europe
ECtHR (Strasbourg)

- ECtHR
- Protocol 12
- ESC

European Union
CJEU (Luxembourg)

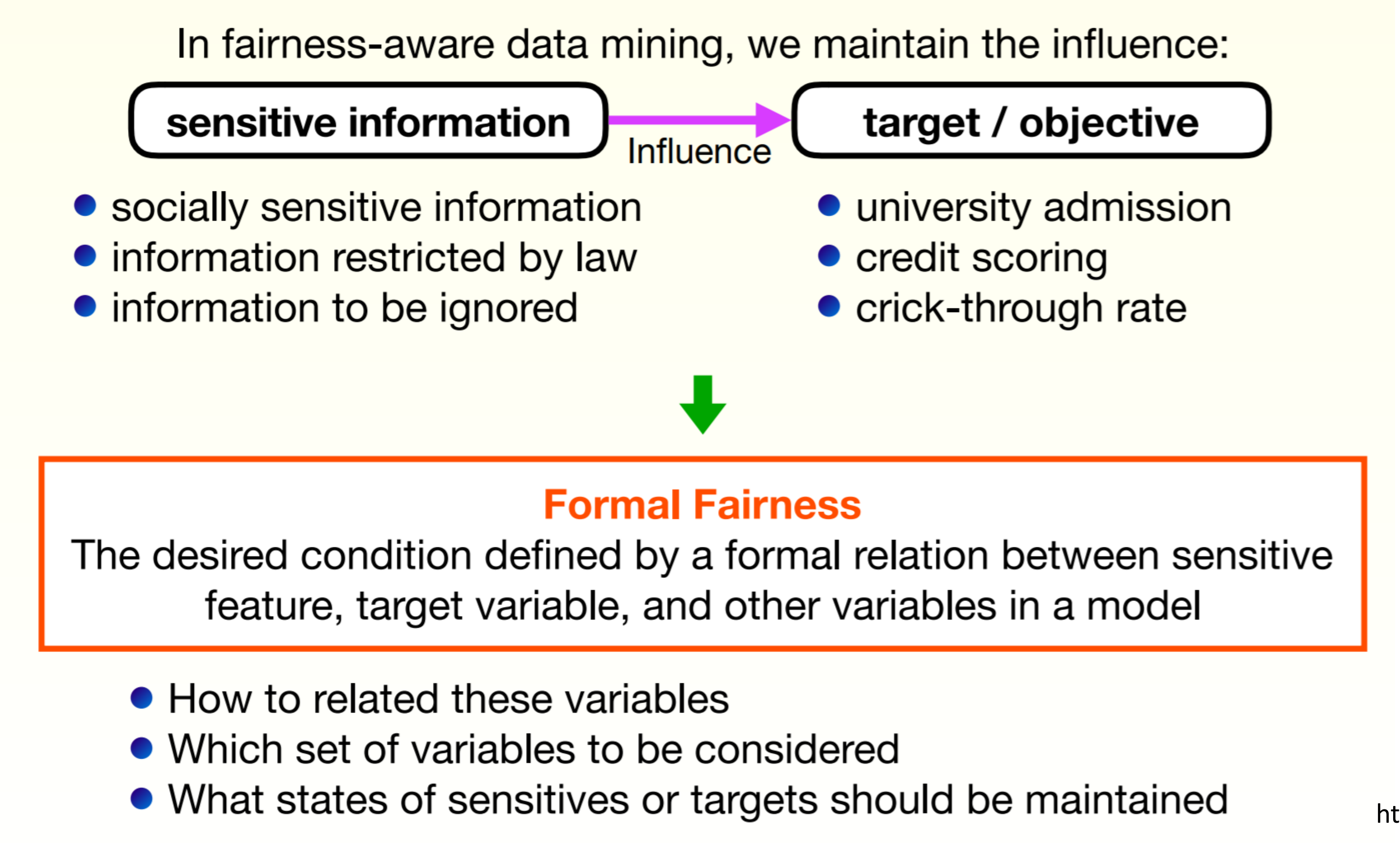
- CRF
- EU Anti-discrimination Law (ie., RED)

EU Member States
National Courts
Consitution

- Criminal Law
- Civil Law
- Social (labour) Law
- Admin Law

https://www.migpolgroup.com/wp-content/uploads/2020/09/MINDSET-Handbook-on-the-Racial-Equality-Directive-003-final.pdf

Formalizing Fairness



In fairness-aware data mining, we maintain the influence:

sensitive information → **target / objective**

Influence

- socially sensitive information
- information restricted by law
- information to be ignored

- university admission
- credit scoring
- crick-through rate

Formal Fairness

The desired condition defined by a formal relation between sensitive feature, target variable, and other variables in a model

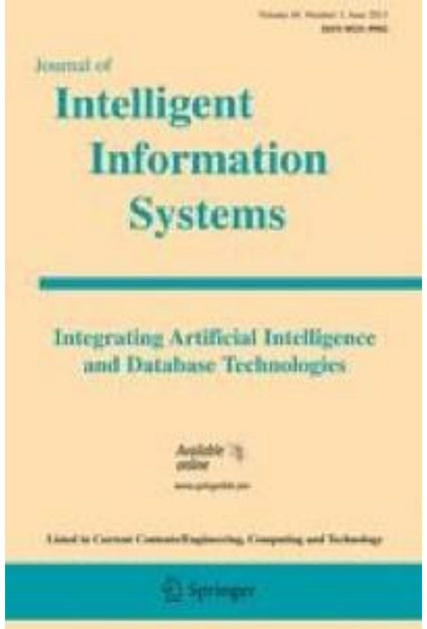
- How to related these variables
- Which set of variables to be considered
- What states of sensitives or targets should be maintained

http://www.kamishima.net/faml/

Fairness, Accountability, and Transparency (Part II)

The Use of Algorithms

- Algorithms silently structure our lives
- Algorithms make data sets valuable
- Algorithms can affect lives in ways far beyond our choice of nightly entertainment



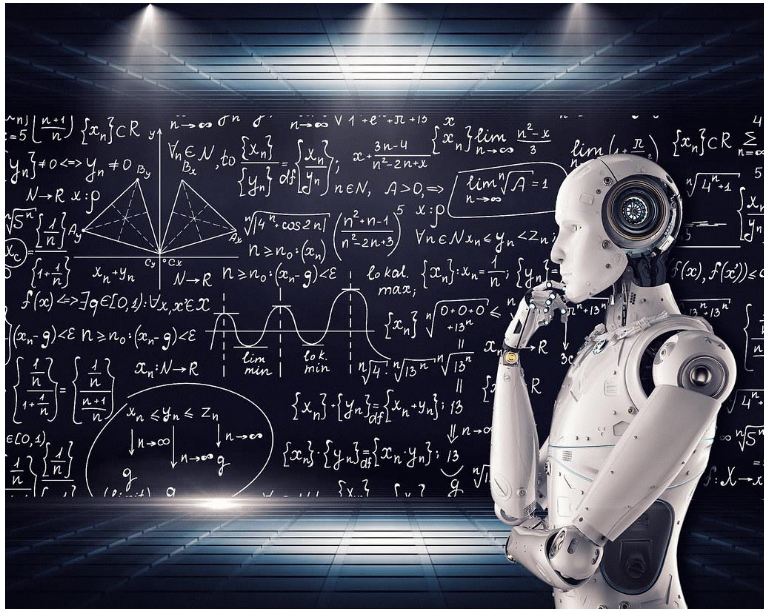
Journal of
Intelligent
Information
Systems

Integrating Artificial Intelligence
and Database Technologies

Available
online

Springer

This Photo by Unknown Author is licensed under CC BY



This Photo by Unknown Author is licensed under CC BY

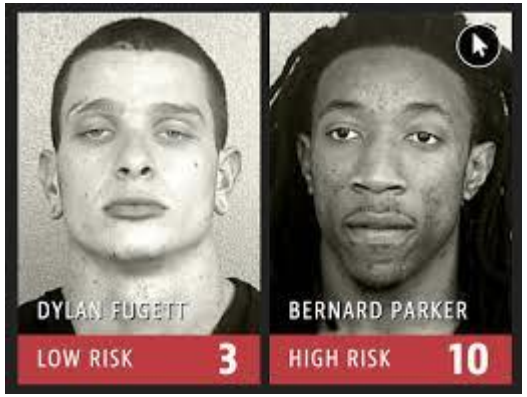
“Algorithms silently structure our lives”

The COMPAS Tool

“Rodríguez was just 16 at the time of his arrest, and was convicted of 2nd degree murder for his role in an armed robbery of a car dealership that left an employee dead. Now, twenty-six years later, he was a model of rehabilitation. He had requested a transfer to Eastern, a maximum-security prison, in order to take college classes. He had spent four and a half years training service dogs for wounded veterans and eleven volunteering for a youth program. A job and a place to stay were waiting for him outside. And he had not had a single disciplinary infraction for the past decade…

Yet, last July, the parole board hit him with a denial. It might have turned out differently but, the board explained, a computer system called COMPAS had ranked him “high risk.” Neither he nor the board had any idea how this risk score was calculated; Northpointe, the for-profit company that sells COMPAS, considers that 2018) **information to be a trade secret.”**

Eastern Correctional Facility in 2015 Source: wikipedia



Source: https://link.springer.com/article/10.1007/s10551-018-3921-3

Algorithms, Decisions, and Accountability

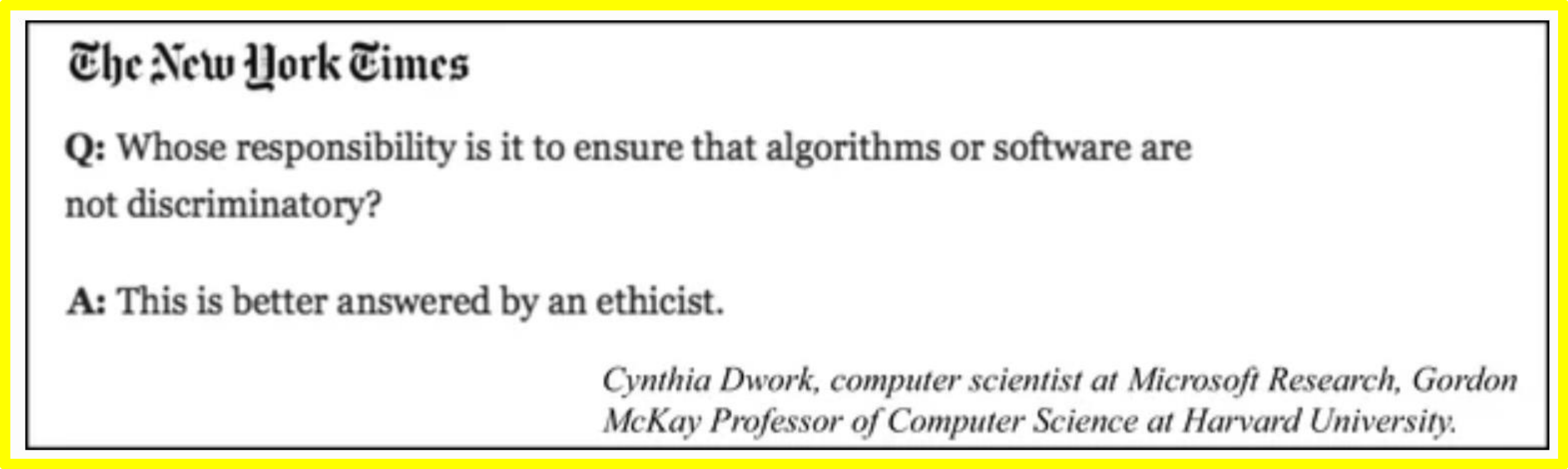
- **Computers are expected to provide unbiased calculations** insofar they take in points of reference objectively and use algorithmic processes to provide a standard outcome

- The conventional wisdom therefore is that **there are considerable benefits of using algorithms over human processing and decision making**
  - speed, efficiency and consistency
  - and even fairness as there a common misconception that algorithms automatically always result in ‘unbiased’ decisions

- Algorithms are therefore attractive because they promise speed and neutrality in decision making.

Algorithms, Decisions and Accountability

- However **algorithmic decision-making is “black boxed”** --- Frank Pasquale
  - what goes into the computer for processing and what the outcome is
  - inner process remains unknown



The New York Times

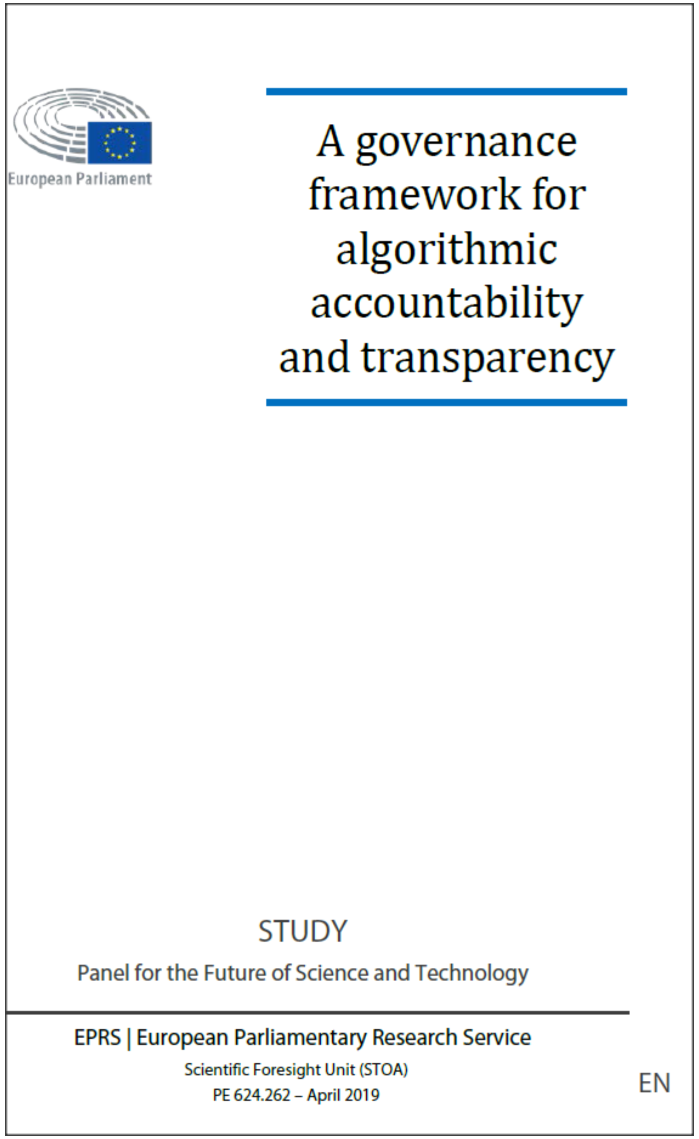
Q: Whose responsibility is it to ensure that algorithms or software are not discriminatory?

A: This is better answered by an ethicist.

Cynthia Dwork, computer scientist at Microsoft Research, Gordon McKay Professor of Computer Science at Harvard University.

Defining Accountability

Accountability - 'A set of mechanisms, practices and attributes that sum to a governance structure which involves committing to legal and ethical obligations, policies, procedures and mechanism, explaining and demonstrating ethical implementation to internal and external stakeholders and remedying any failure to act properly’.
Source: EPRS (2019)



European Parliament

A governance framework for algorithmic accountability and transparency

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)
PE 624.262 – April 2019

EN

Meaning And Functions Of Accountability

Accountability is primarily a **legal** and ethical obligation on an individual or organisation to account for its activities, accept **responsibility** for them, and to disclose the results in a transparent manner

Functions of accountability principle

- act as a deterrent to reckless, irresponsible or illegal behaviour on the part of humans deploying/using algorithmic systems
- generate a self-reflective feedback loop for citizens and society, exposing existing biases and power dynamics

Algorithmic Accountability

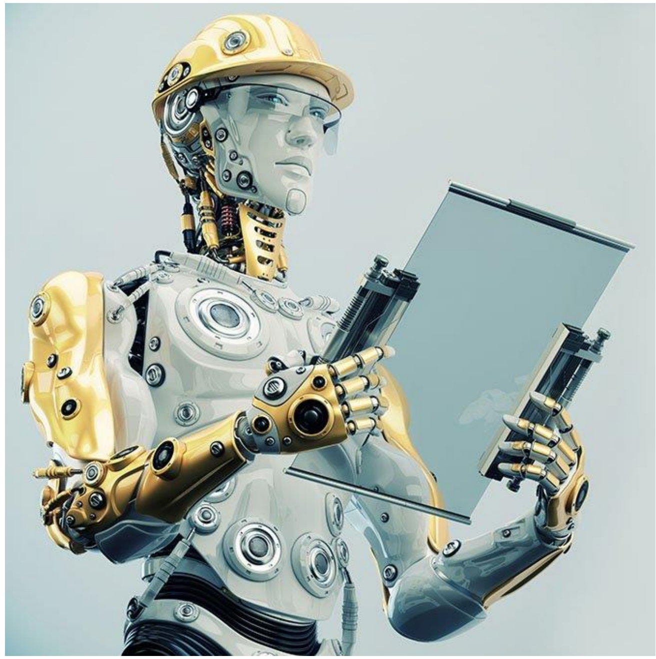
Algorithmic accountability refers to the assignment of responsibility for how an algorithm is created and its impact on society.

The final decisions to put an algorithmic system on the market belongs to the technology's designers and company. ***Critically, algorithms do not make mistakes, humans do***

Assigning responsibility is critical for quickly remediating discrimination and assuring the public that proper oversight is in place.

In addition to clearly assigning responsibility, **accountability must be grounded in enforceable policies** that begin with auditing in pre- and post- marketing trials as well as standardized assessments for any potential harms.

Accountability basic function is to act as a deterrent to reckless, irresponsible or illegal behaviour on the part of humans deploying/using algorithmic systems.

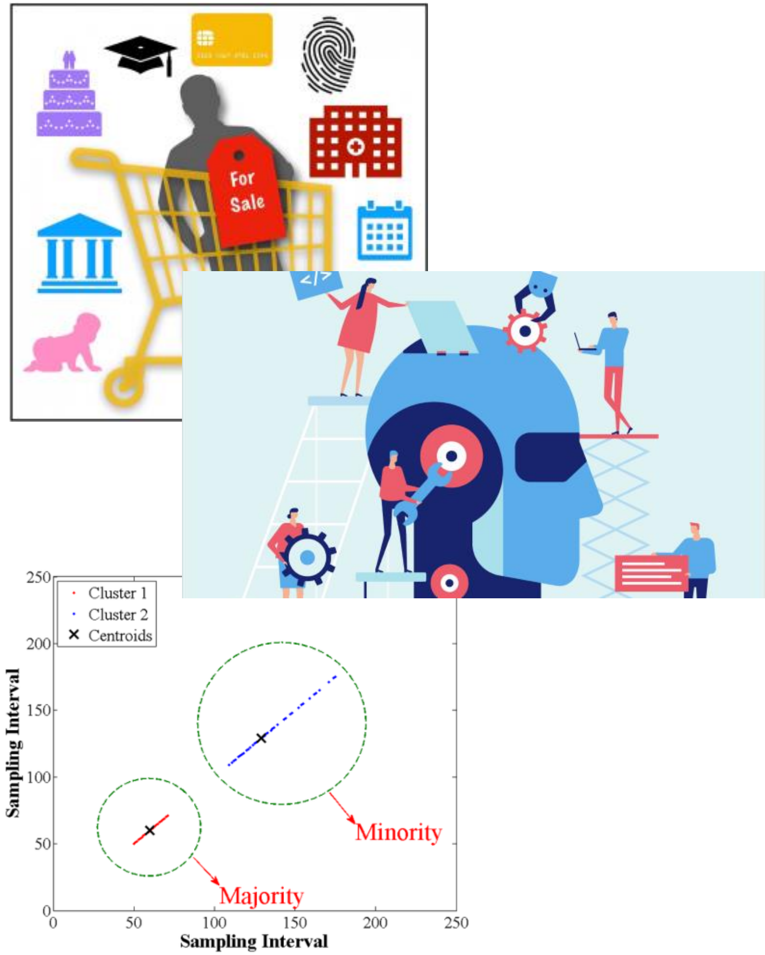


Challenges For Algorithmic Accountability (1/2)

1. **Complex interactions between sub-systems and data sources,** some of which might not be under the control of the same entity.

2. **Unexpected outcomes associated with the impossibility of testing** against all possible input conditions when there are no methods for generating formal proofs for the system’s performance.

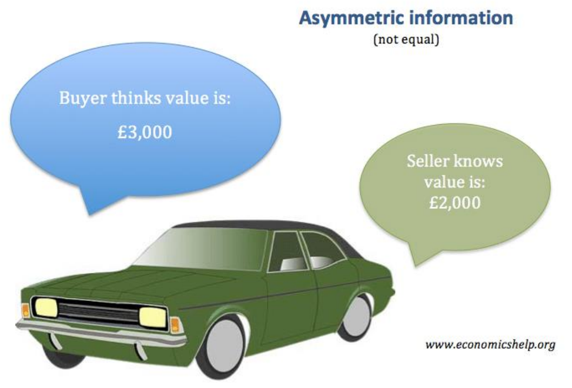
3. **Difficulties in translating algorithmically derived concepts into human understandable concepts resulting in incorrect interpretations** of the meaning of algorithmic results.



Src: Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). A governance framework for algorithmic accountability and transparency.

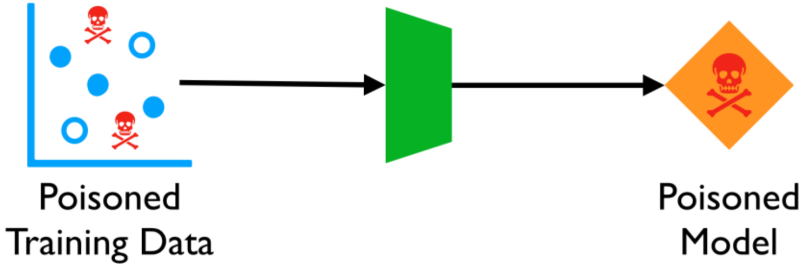
Challenges For Algorithmic Accountability (2/2)

4. **Information asymmetries arising from algorithmic inferences and black box processes**



5. **Ubiquity of (small) algorithmic decisions** which, if systematically biased, may accumulate to have significant impacts on people

6. **Purposeful injections of adversarial data** to fool a system into making errors, often in ways that can be very difficult to detect.



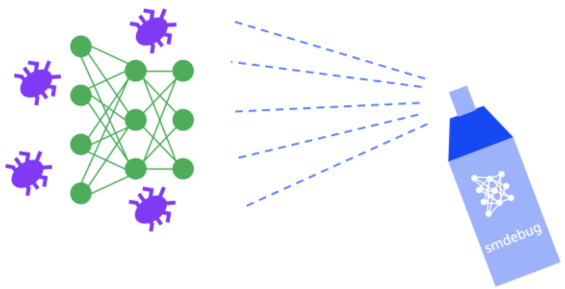
Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (2019). A governance framework for algorithmic accountability and transparency.

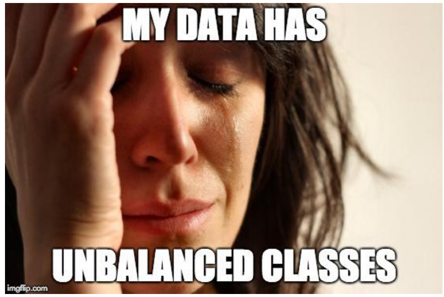
Algorithmic Accountability

Even if algorithms are programmed with specific attention to well-defined legal norms, it is difficult to know whether the algorithm behaved according to the legal standard or not, in any given circumstance.

Algorithms that engage in discrimination offer a good example for this point.

Tracing the discrimination to a problem with the algorithm could be nearly impossible





Algorithmic Accountability

Accountability for actions taken by algorithmic systems may need to be different than for human actions, those differences are largely governed by the particular application.

As a result, we will only look at mechanisms for ensuring that algorithmic systems satisfy specifications

Process standards and certification, such as ISO/IC JTC 1/SC7 standards for software engineering, or the Capability Maturity Model Integration (processes and procedures organisations should follow in systems design)

The IEEE P7000 series addresses specific issues at the intersection of technological and ethical considerations (particularly IEEE P7001 and P7003)

The most recent one is IEEE 7000™-2021: Model Process for Addressing Ethical Concerns During System Design, published in Sep 2021

Principal Of Accountability Is Contested

"Currently, it is difficult to get technology corporations to answer for the harms their products have caused." (Caplan et al., 2018)

Why should this be?

> **Developers argue** that algorithms are objective, neutral, blank-slates leaving minimal responsibility for the developer. It is the responsibility of users to interpret the results correctly.
>
> **Users on the other hand argue** that algorithms are black boxes, complicated, difficult to explain.

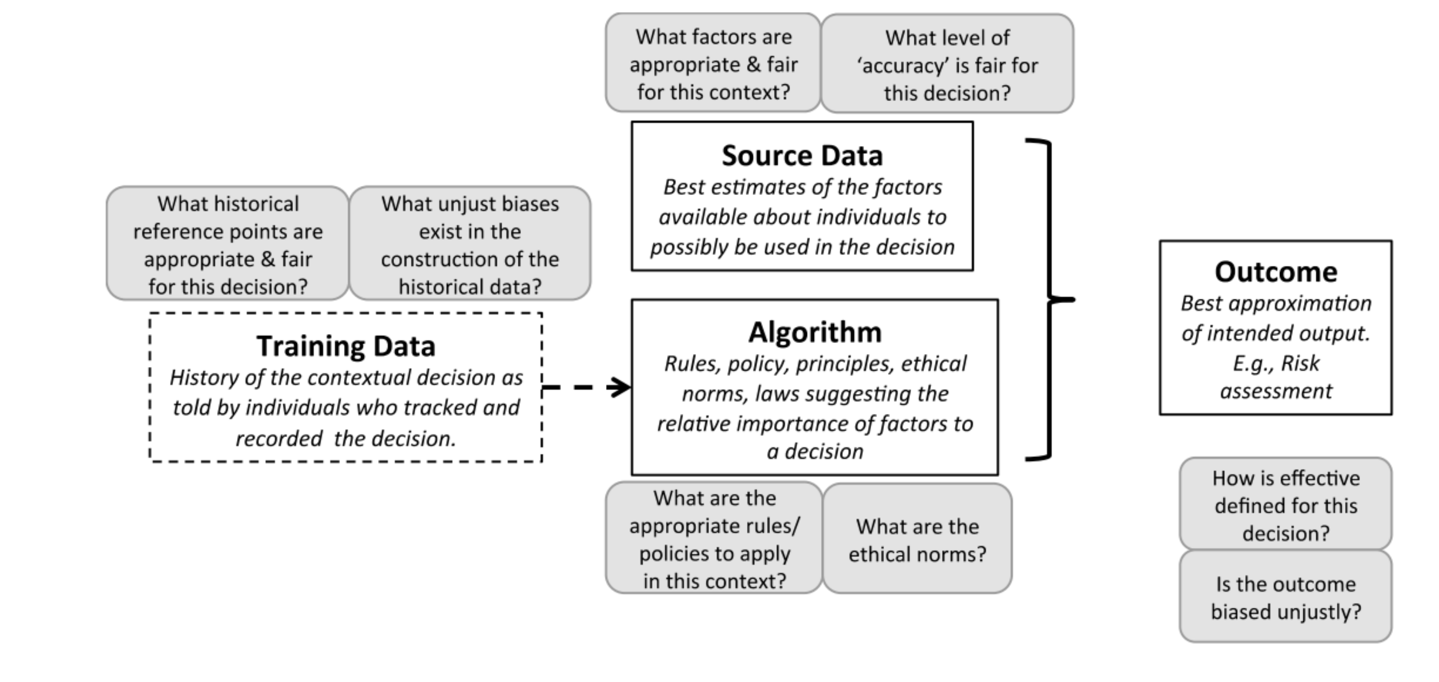
Algorithms have life-changing implications and therefore moral consequences… Who should be accountable for them particularly for redress if they go wrong?

Need For Accountable Development

Algorithms are viewed as maximizing efficiency or accuracy; computer scientists are, therefore, responsible for ensuring efficiency and accuracy

Algorithms are not neutral but are "value-laden" to pre-determine roles and responsibilities to make choices within the algorithmic system

These questions go unanswered leaving the development process ultimately unaccountable



What factors are appropriate & fair for this context?

What level of 'accuracy' is fair for this decision?

Source Data
Best estimates of the factors available about individuals to possibly be used in the decision

What historical reference points are appropriate & fair for this decision?

What unjust biases exist in the construction of the historical data?

Training Data
History of the contextual decision as told by individuals who tracked and recorded the decision.

Algorithm
Rules, policy, principles, ethical norms, laws suggesting the relative importance of factors to a decision

Outcome
Best approximation of intended output. E.g., Risk assessment

What are the appropriate rules/ policies to apply in this context?

What are the ethical norms?

How is effective defined for this decision?

Is the outcome biased unjustly?

Fig. 3 Adding in missing masses to algorithm decision-making process

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, *160*(4), 835-850.

Framework For Appropriate Level Corporate Accountability

Martin (2018) proposes that the level of corporate accountability depends on two decisions that contribute to the appropriate type of accountability expected

- ○**Role of the algorithm in a decision** (e.g is it deciding to target an advertisement at us or is it being using as the primary basis for granting us a mortgage?) **[y-axis]**
- ○**Significance of the decision in societal terms** (e.g., is it suggesting a film to watch or recommending whether we should get released from prison?) **[x-axis]**

A range of circumstances will determine what the responsibility and accountability of a firm who developed the algorithm. This will depend upon (1)whether the algorithm plays a large or small role in the decision (2) the context in which the algorithm is to be used .

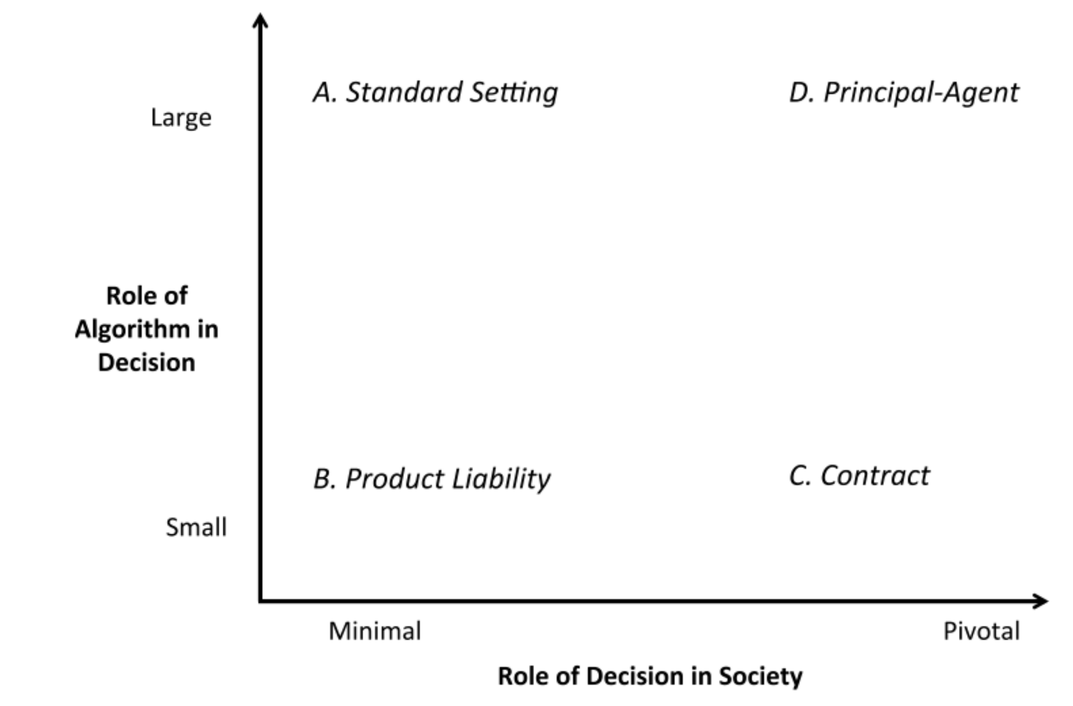


Fig. 4 Firm responsibility for algorithms

Algorithmic Impact Assessments -AIA-(1/3)

Algorithmic Impact Assessments provide a framework designed to help policymakers and their constituents understand where algorithmic systems are used within government, assess the intended use and proposed implementation, and allow community members and researchers to raise concerns that require mitigation.

The general shape of the process is likely to include the following:

❑**Publication** of public authority’s definition of 'algorithmic system'. This allows the public to understand how the authority decides which systems will be subjected to AIAs.

❑Once the definition has been published and gone **through public review**, it is used to assess all currently used systems

Algorithmic Impact Assessments -AIA-(2/3)

❑Public disclosure of purpose, scope, intended use and associated policies/practices, self-assessment timeline/process and potential implementation timeline of the algorithmic system OR publication (and archiving) of the decision not to review a potential system

❑Publication of plan for meaningful, ongoing access to external researchers to review the system once it is deployed

❑Public participation period

❑Publication of final Algorithmic Impact Assessment, once issues raise in public participation have been addressed

❑Renewal of AIAs on a regular timeline

❑Opportunity for public to challenge failure to mitigate issues raised in the public participation period or foreseeable outcomes

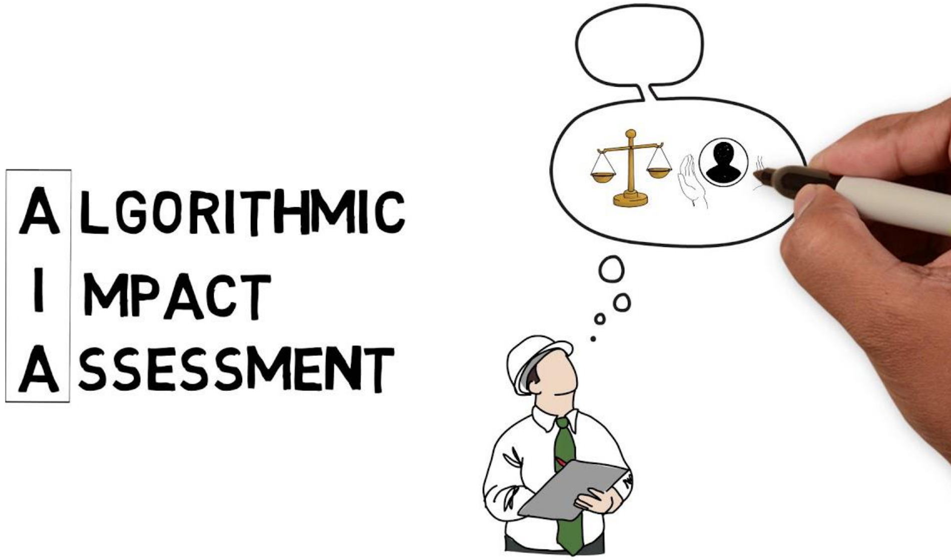
Algorithmic Impact Assessments -AIA-(3/3)

1.Capture an AI system’s risk. Establishing “risk gating criteria” enables organizations to properly classify the level of scrutiny needed for a specific AI application.

2.Cover full development life-cycle requirements. An AIA should encompass strategy and planning, ecosystem analysis, implications to model development, issues related to training data, deployment and, finally, ongoing operation, monitoring issues and governance.

3.Assess impact and increase accountability through a multi-stakeholder analysis. A successful impact assessment engages a broad range of internal stakeholders and may also include external representatives, such as ethics or data review boards.

4.Facilitate go/no-go decisions. This goal should address whether a model should move to production, determine if it’s ready to be transitioned for business-as-usual operations and decide whether it should continue as-is or be retrained, redesigned or retired.



https://www.pwc.com/us/en/tech-effect/ai-analytics/algorithmic-impact-assessments.html

Principles For Accountable Algorithms

*Algorithms and the data that drive them are designed and created by people -- There is always a human ultimately responsible for decisions made or informed by an algorithm. "**The algorithm did it" is not an acceptable excuse** if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes.*

Fairness, Accountability, and Transparency in Machine Learning

Responsibility

Make available **externally visible avenues of redress** for adverse individual or societal effects of an algorithmic decision system, and **designate an internal role** for the person who is responsible for the timely remedy of such issues.

Explainability

Ensure that algorithmic decisions as well as any data driving those decisions **can be explained to end-users** and other stakeholders in non-technical terms.

Accuracy

Identify, log, and articulate **sources of error and uncertainty** throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures.

Auditability

Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.

Fairness

Ensure that algorithmic decisions **do not create discriminatory or unjust impacts** when comparing across different demographics (e.g. race, sex, etc).

Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML

Social Impact Statement for Algorithms

Algorithm creators should develop a Social Impact Statement using the principles as a guiding structure.
This statement should be revisited and reassessed (at least) three times during the design and development process:

- design stage,
- pre-launch,
- and post-launch.

The Social Impact Statement should minimally answer the questions of each principle. It is also included some concrete steps that can be taken, and documented as part of the statement, to address these questions. These questions and steps make up an outline of such a social impact statement.

GDPR And Algorithm Accountability

From GDPR Recital 71 (Profiling): "In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, **the controller should use appropriate mathematical or statistical procedures for the profiling**, implement technical and organisational **measures** appropriate **to ensure**, in particular, that factors which result in **inaccuracies in personal data are corrected and the risk of errors is minimised**, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and **that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation**, or that result in measures having such an effect."

GDPR And Algorithm Accountability

- Recall that a data subject has a right to contest an individual algorithmic decision (GDPR Art.22), to receive notice of solely automated decision-making (GDPR Art.13), and to request access to "meaningful information about the logic involved" (GDPR Art.15). In this way, data controllers can be challenged and thus accountable for algorithmic decisions.
- GDPR Article 35 (Data protection impact assessment) further provides that where data processing (in particular using new technologies) is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data. A single assessment may address a set of similar processing operations that present similar high risks and that the controller consult with the data protection officer when carrying out a **data protection impact assessment**.
- It is argued (Kaminski and Malgieri, 2019) that this requirement together with the recital, provides a case that "GDPR's version of an Algorithmic Impact Assessment (that is DPIA) serves as a central connection between its two approaches to regulating algorithms: individual rights and systemic governance."

Kaminski, M. E. and Malgieri, G. (2019) 'Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations', *SSRN Electronic Journal*, pp. 1–29

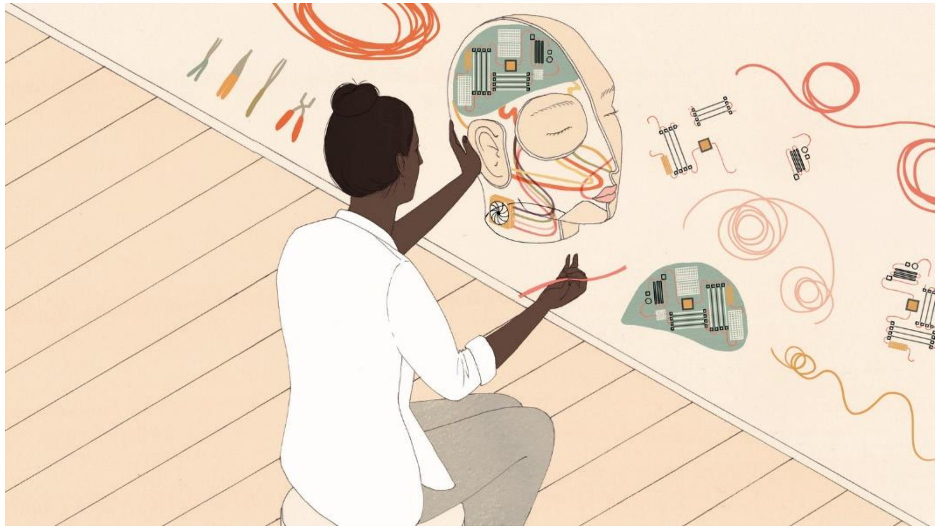
Methods And Tools And Standards For Ensuring That Algorithms Comply With Fairness Policies

How to prevent machine bias

▶**Use a representative dataset**. Feeding your algorithm representative data is THE most important aspect when it comes to preventing bias in machine learning.

▶**Choose the right model**. Every AI algorithm is unique and there is no single model that can be used to avoid bias. ...

▶**Monitor and review**.



IEEE P7003 TM - Algorithmic Bias Considerations

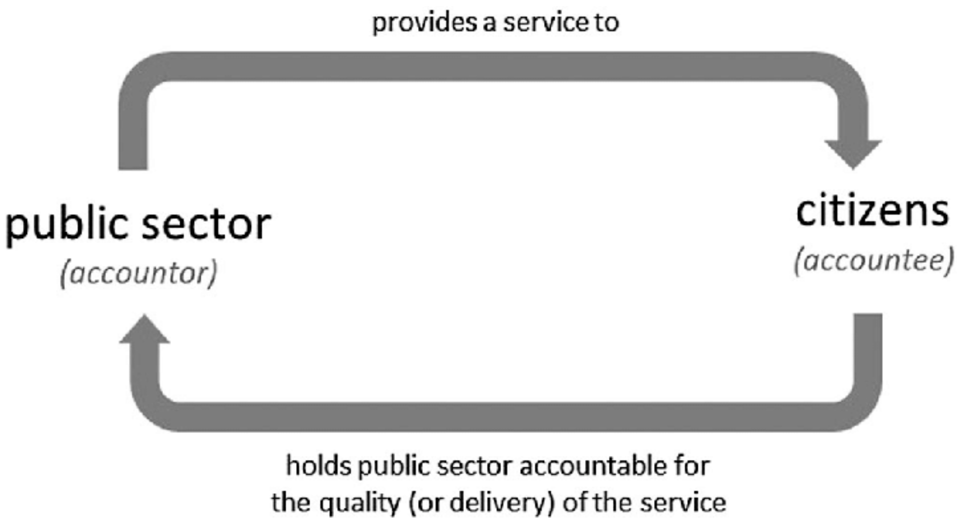
- The IEEE P7003 standard will provide a framework, which helps developers of algorithmic systems and those responsible for their deployment to identify and mitigate unintended, unjustified and/or inappropriate biases in the outcomes of the algorithmic system
- IEEE P7003 will allow algorithm creators to communicate to regulatory authorities and users that the most up-to-date best practices are used in the design, testing and evaluation of algorithms in order to avoid unjustified differential impact on users.
- The standard committed to the support and responsibility of algorithm creators to prioritize accountability and ethics within new frameworks of all levels of design, testing and evaluation, reducing the impact of discrimination and encouraging neutrality and fairness for future technologies
- This standard will provide the required accountability to show algorithms are developed and applied without issues of negative bias aimed at protected characteristics of individuals or groups, including such considerations as race, gender and sexuality.

Accountability in public sector use of algorithmic decision making

- Algorithmic systems are increasing being used by public authorities to improve efficiencies, implement complex processes and support evidence-based policy making.

- Due to the nature of public sector responsibilities, these uses of algorithmic systems have potentially far-reaching impacts sometimes involving the weakest members of society.

The use of algorithmic systems in public services therefore requires extra levels of transparency and accountability.


provides a service to
public sector
(accountor)
citizens
(accountee)
holds public sector accountable for
the quality (or delivery) of the service

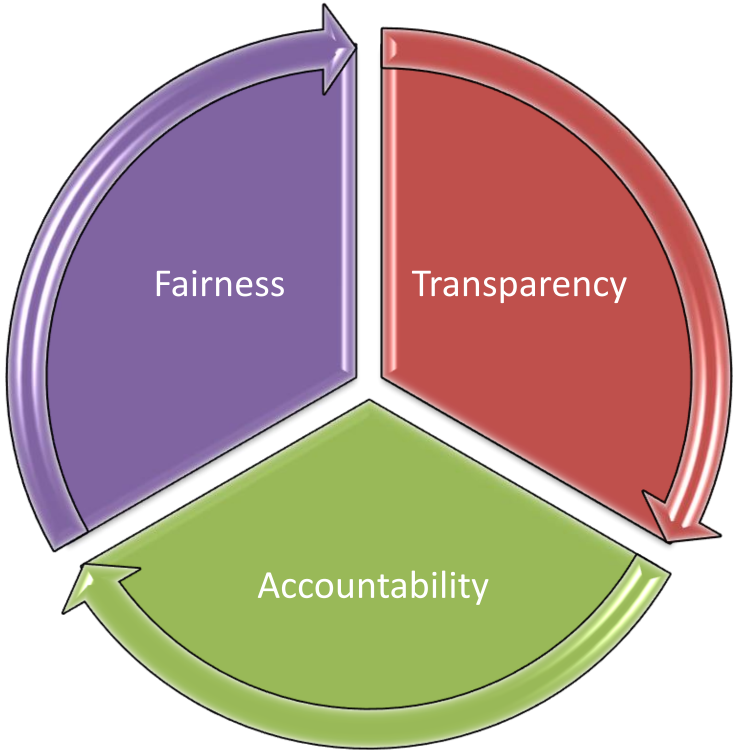
Fairness, Accountability, and Transparency (Part III)

Recap – FAT in ML

Fairness, Accountability and Transparency (FAT) – a way of examining critically **one of the principal concerns about the use of algorithms in society** (as part of the broader socio-technical system)

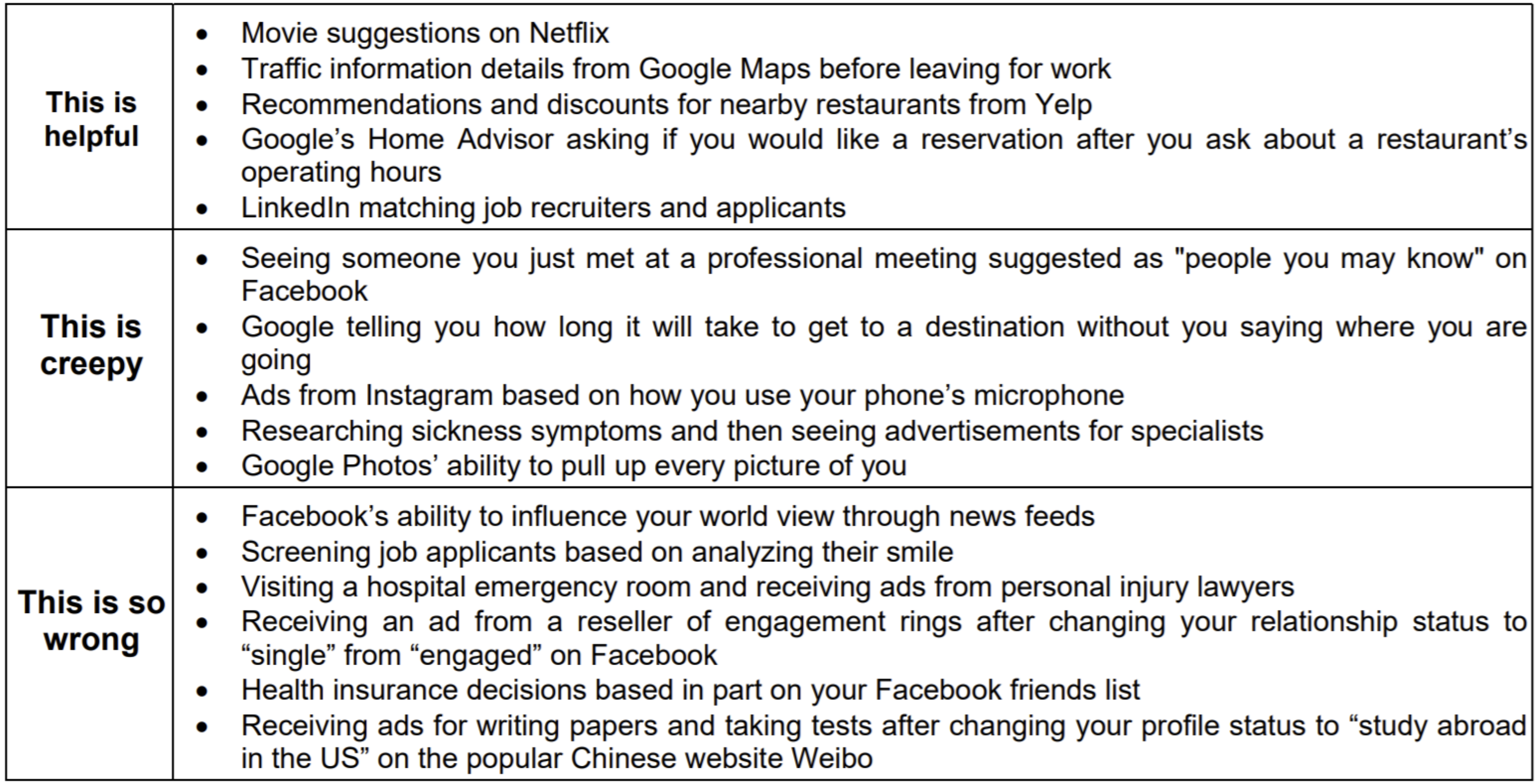
“In principle, **fairness** is the absence of any bias based on an individual’s inherent or acquired characteristics that are irrelevant in the particular context of decision-making.”

To assess fairness, we need transparency
To enforce accountability, we need transparency



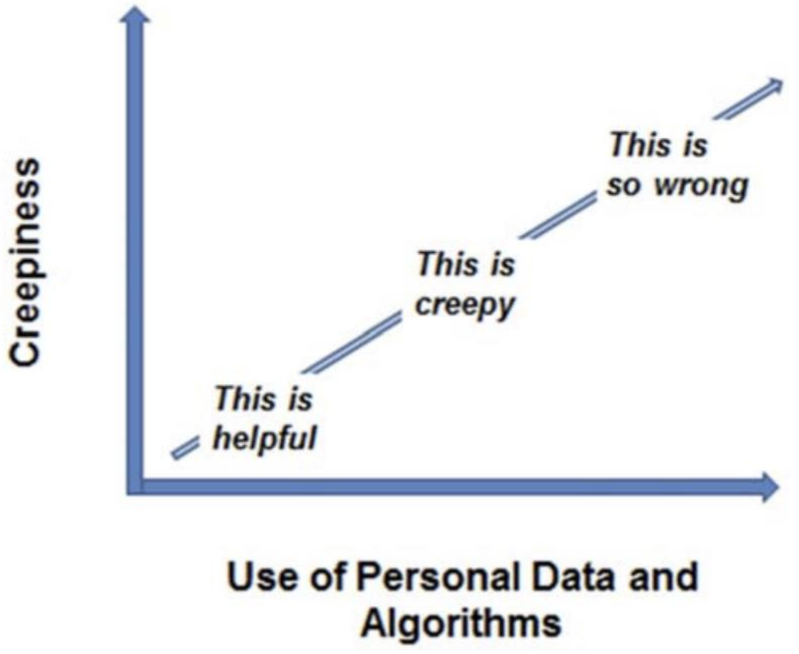
Algorithmic Transparency

Creepiness Scale: People’s reactions to recommendations based on personal data and algorithms



This is helpful	• Movie suggestions on Netflix • Traffic information details from Google Maps before leaving for work • Recommendations and discounts for nearby restaurants from Yelp • Google’s Home Advisor asking if you would like a reservation after you ask about a restaurant’s operating hours • LinkedIn matching job recruiters and applicants
This is creepy	• Seeing someone you just met at a professional meeting suggested as "people you may know" on Facebook • Google telling you how long it will take to get to a destination without you saying where you are going • Ads from Instagram based on how you use your phone’s microphone • Researching sickness symptoms and then seeing advertisements for specialists • Google Photos’ ability to pull up every picture of you
This is so wrong	• Facebook’s ability to influence your world view through news feeds • Screening job applicants based on analyzing their smile • Visiting a hospital emergency room and receiving ads from personal injury lawyers • Receiving an ad from a reseller of engagement rings after changing your relationship status to “single” from “engaged” on Facebook • Health insurance decisions based in part on your Facebook friends list • Receiving ads for writing papers and taking tests after changing your profile status to “study abroad in the US” on the popular Chinese website Weibo

The Creepiness Scale for the Use of Personal Data and Algorithms

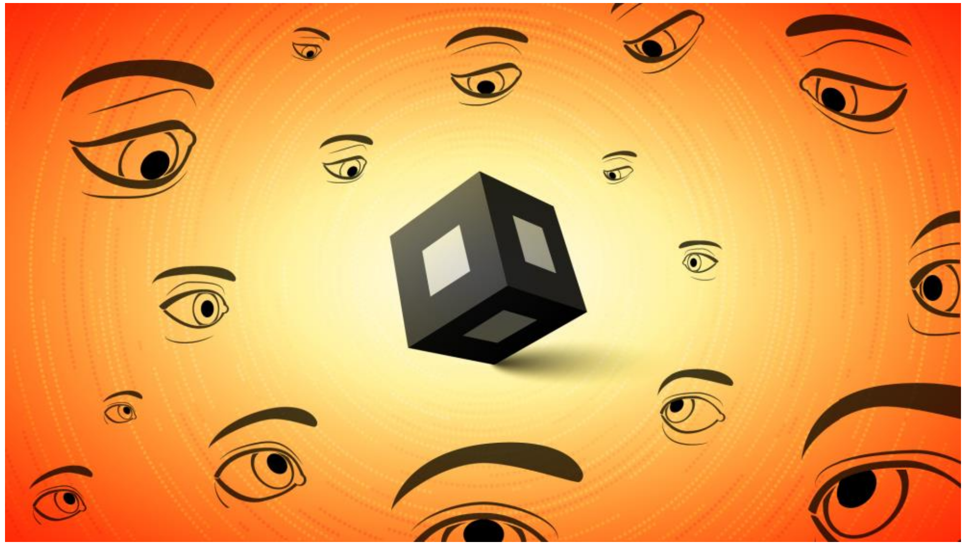


Watson, H. J., & Nations, C. (2019). Addressing the Growing Need for Algorithmic Transparency. *Communications of the Association for Information Systems*, *45*(1), 26.

Algorithmic Transparency

- If an algorithm can be defined as **a set of steps** that a computer program follows in order to make a decision about a particular course of action, then

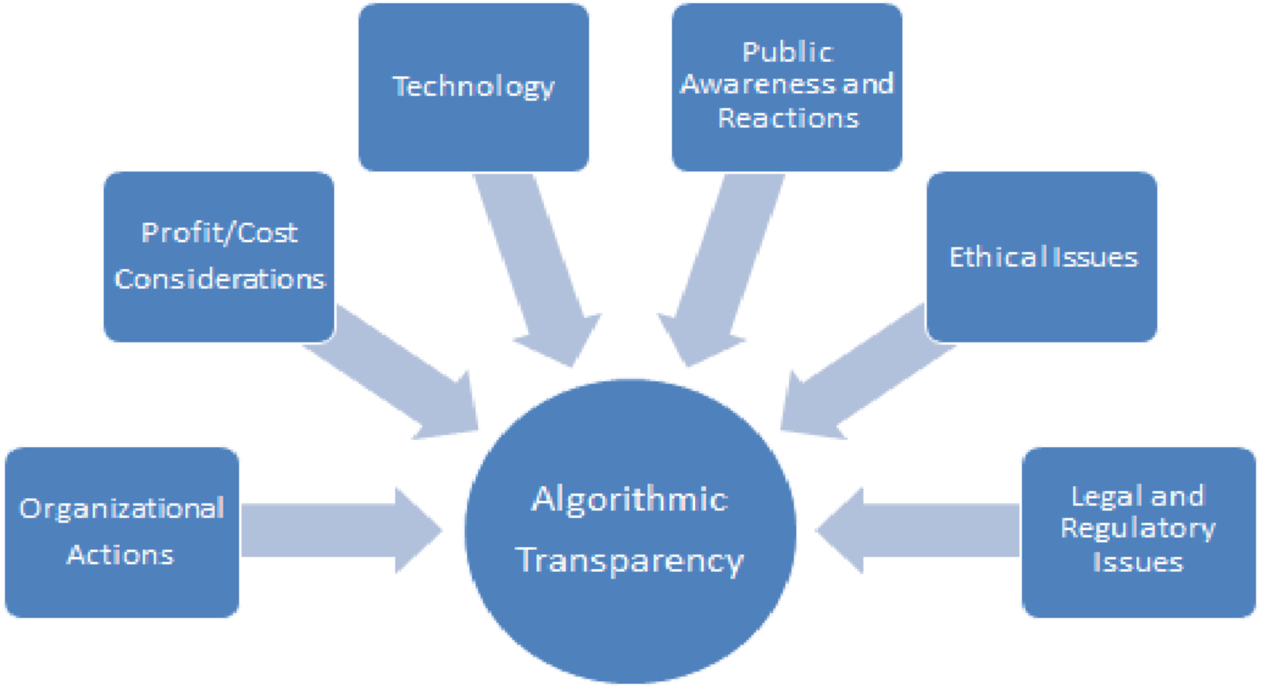
- its **transparency refers to the degree of openness about these steps, their purpose, structure and underlying actions of the algorithms** that are used to search for, process and deliver information.



Factors that Affect Algorithmic Transparency

Multiple factors influence algorithmic transparency

With care, companies can benefit from and avoid the penalties associated with using personal data and algorithms inappropriately

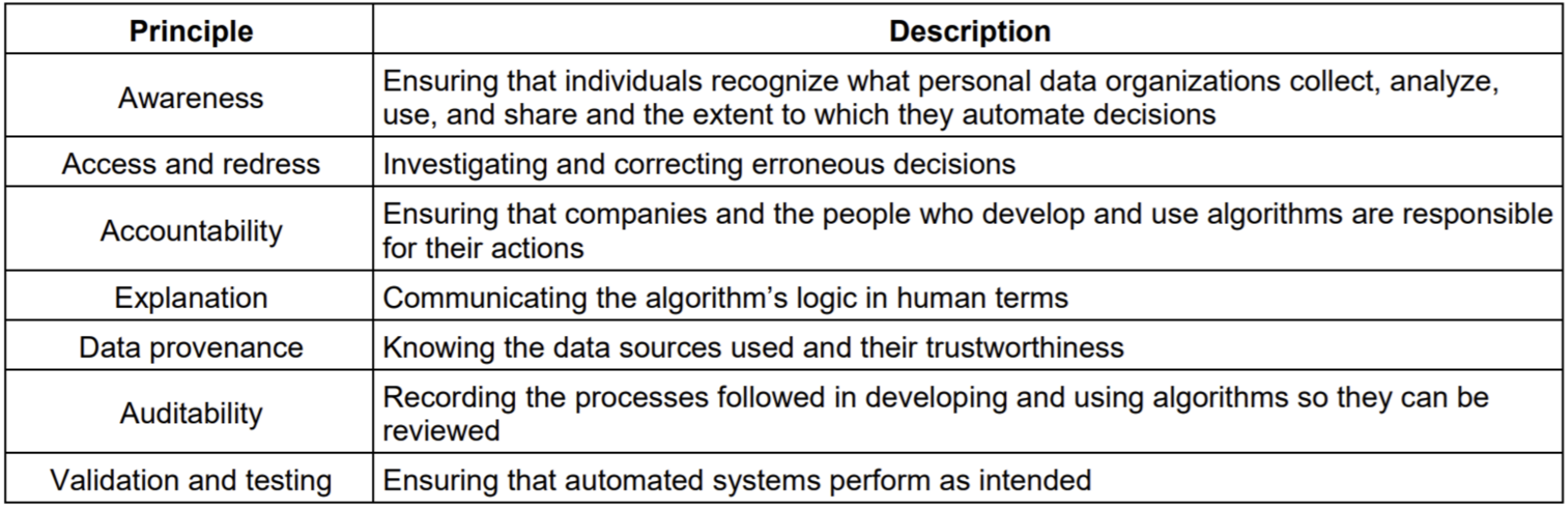


Technology
Public Awareness and Reactions
Profit/Cost Considerations
Ethical Issues
Organizational Actions
Algorithmic Transparency
Legal and Regulatory Issues

Factors that Affect Algorithmic Transparency

Algorithmic Transparency Principles

ACM principles for ensuring that organizations use personal data and algorithms fairly.
Aim is to minimize potential harms while realizing the benefits of algorithmic decision-making

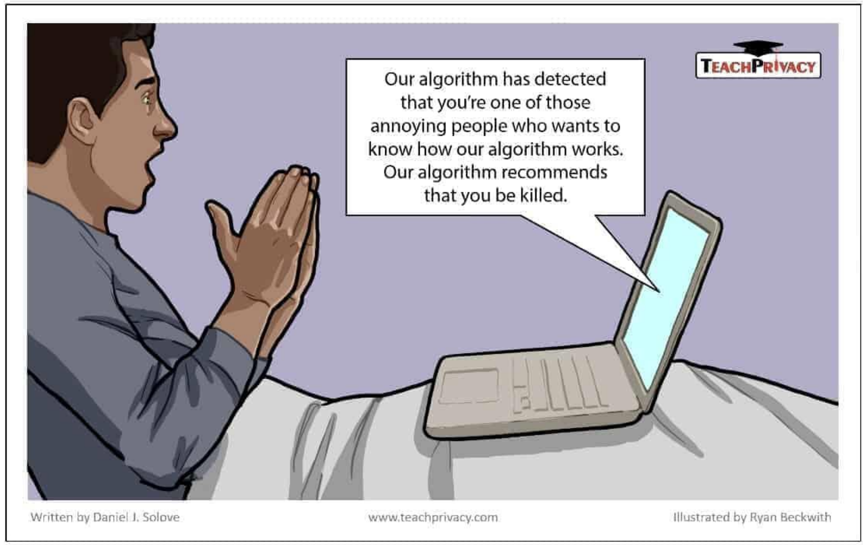


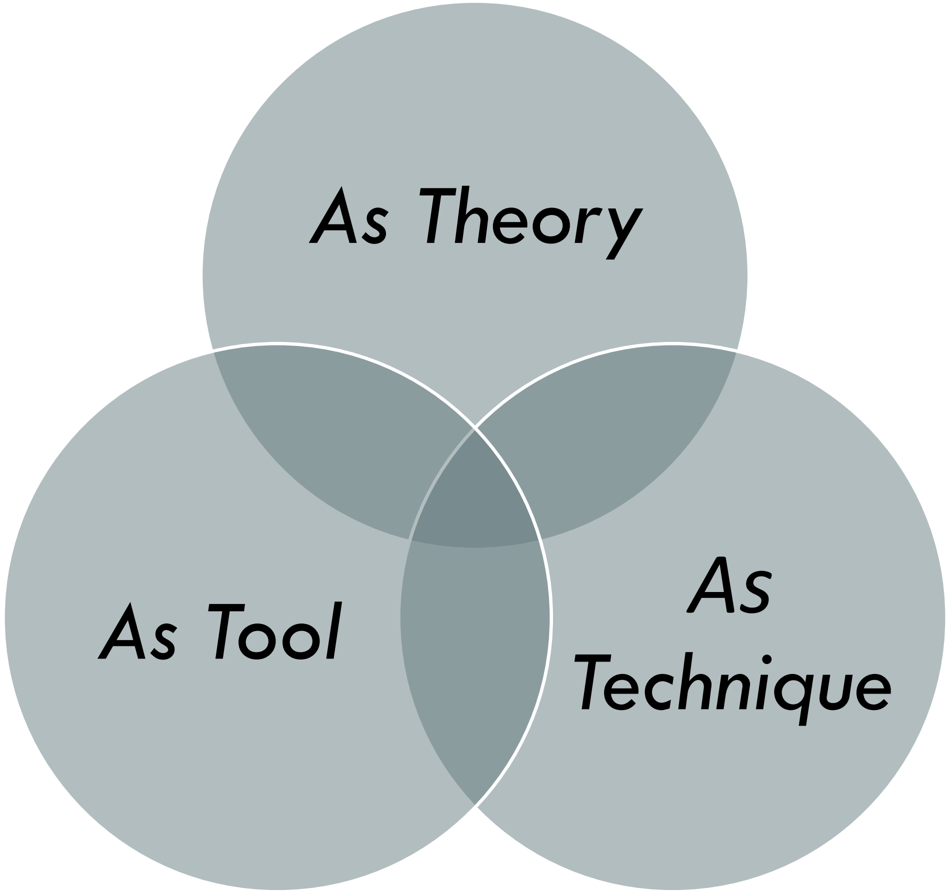
Principle	Description
Awareness	Ensuring that individuals recognize what personal data organizations collect, analyze, use, and share and the extent to which they automate decisions
Access and redress	Investigating and correcting erroneous decisions
Accountability	Ensuring that companies and the people who develop and use algorithms are responsible for their actions
Explanation	Communicating the algorithm’s logic in human terms
Data provenance	Knowing the data sources used and their trustworthiness
Auditability	Recording the processes followed in developing and using algorithms so they can be reviewed
Validation and testing	Ensuring that automated systems perform as intended

Principles for Algorithmic Transparency and Accountability

The Uses of Transparency

- Theory – relationship to fairness and accountability; how it underpins both
- Tool – who, what, why?
- Technique – practical methods to implement


Our algorithm has detected that you're one of those annoying people who wants to know how our algorithm works. Our algorithm recommends that you be killed.
TeachPrivacy
Written by Daniel J. Solove
www.teachprivacy.com
Illustrated by Ryan Beckwith



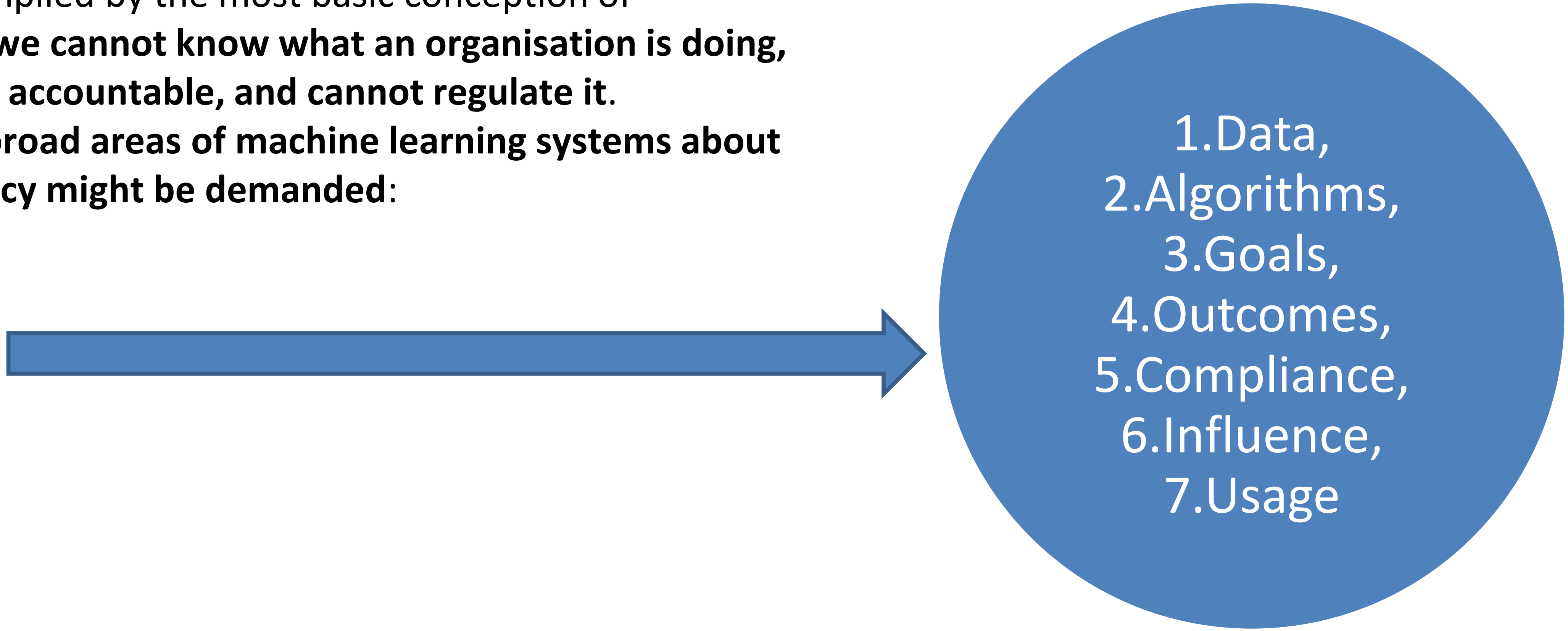
Transparency as a tool to answer a range of questions (what, how, who?)

- **Transparency is a tool to uncover the fairness properties of algorithms**
- Tools have three dimensions that are directly relevant to the use of transparency as a mechanism of governance of algorithmic systems:

1. **What?** A tool is valuable not in itself but because **of the goals its serves**; a can-opener is only useful if there are cans to be opened.
2. **Why/purpose?** No tool is right for every job. Misusing a tool has costs. Even using it appropriately often requires trade-offs.
3. **Who/whom?** the motivation of the person who **uses the tool** and what their ultimate objective is can be an important factor in its effectiveness.

Transparency of what?

Transparency is implied by the most basic conception of accountability: **if we cannot know what an organisation is doing, we cannot hold it accountable, and cannot regulate it**.
There are **seven broad areas of machine learning systems about which transparency might be demanded**:



1. Data,
2. Algorithms,
3. Goals,
4. Outcomes,
5. Compliance,
6. Influence,
7. Usage

Source: European Parliamentary Research Service (2019)

Transparency of what? (1. Data)

1. Data

The transparency of the data used by the algorithmic system -- in particular by machine learning and deep learning algorithms -- can refer to knowledge of

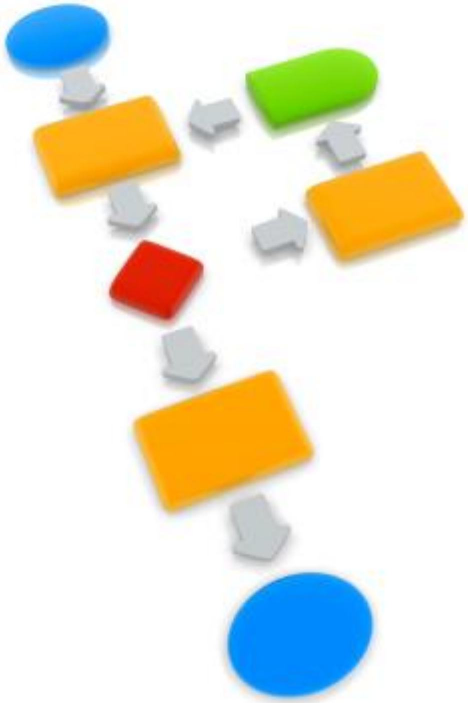
- the raw data,
- the data’s sources,
- how the data were preprocessed,
- the methods by which it was verified as unbiased and representative (including looking for features that are proxies for information about protected classes),
- the processes by which the data are updated and the system is trained on them.

Transparency of what? (2. Algorithms)

2. Algorithms

The transparency of the systems' algorithms can refer to a number of approaches

- ❖testing its output against inputs for which we know the proper output,
- ❖reducing the variables to the most significant so we can validate them,
- ❖testing the system with counterfactuals to see if prejudicial data is infecting the output,
- ❖a third party code review, analysis of how the algorithms work, inspection of internal and external bug reports, or assurance the software development processes are sound.



Transparency of what? (3. Goals)

3. Goals

Transparency of Algorithmic systems can also refer to its goal or goals.

- ❖What is it aiming to do?
- ❖When a system has multiple goals, this would mean being transparent about their relative priorities.
- ❖For example, the AI driving autonomous vehicles (AVs) might be aimed at reducing traffic fatalities, lowering the AVs’ environmental impact, reducing serious injuries, shortening transit times, avoiding property damage, and providing a comfortable ride. A manufacturer could be required to be transparent about those goals and their priority.

Transparency of what? (4. Outcomes)

4. Outcomes

Manufacturers or operators of algorithmic systems could be required to be transparent about

- the outcomes of the deployment of their algorithmic systems, including the internal states of the system (how worn are the brakes of an AV? how much electricity used?),
- the effects on external systems (how many accidents, or times it’s caused another AV to swerve?), and
- computer-based interactions with other algorithmic systems (what communications with other AVs, what data fed into traffic monitoring systems?).

Transparency of what? (5. Compliance)

5. Compliance

Manufacturers or operators may be required to be transparent about their overall compliance with whatever transparency requirements have been imposed upon them.

In many instances, we may insist that these compliance reports are backed by data that is inspectable by regulators or the general public.

Compliance with the transparency of data collection is at the heart of the GDPR. Data subjects need to be informed that their data is being collected and for what purpose.

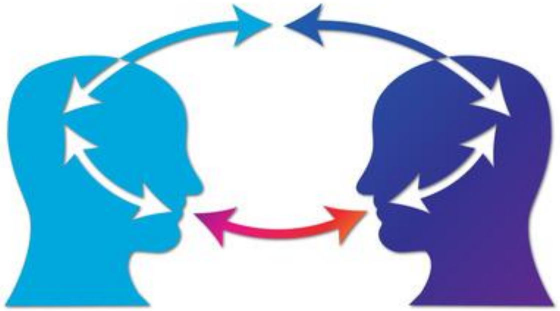
Transparency of what? (6. Influence)

6. Influence

Just as the public has an interest in knowing if an article in a newspaper was in fact paid for by an interested party, the public may have an interest in knowing if any element of the AI process was purposefully bent to favour a particular outcome.

For example, if a trusted search platform is artificially boosting some results because they were paid to, and if it is not flagging that fact to users, users can be manipulated.

Regulators might want to insist that such influence be conspicuously acknowledged.

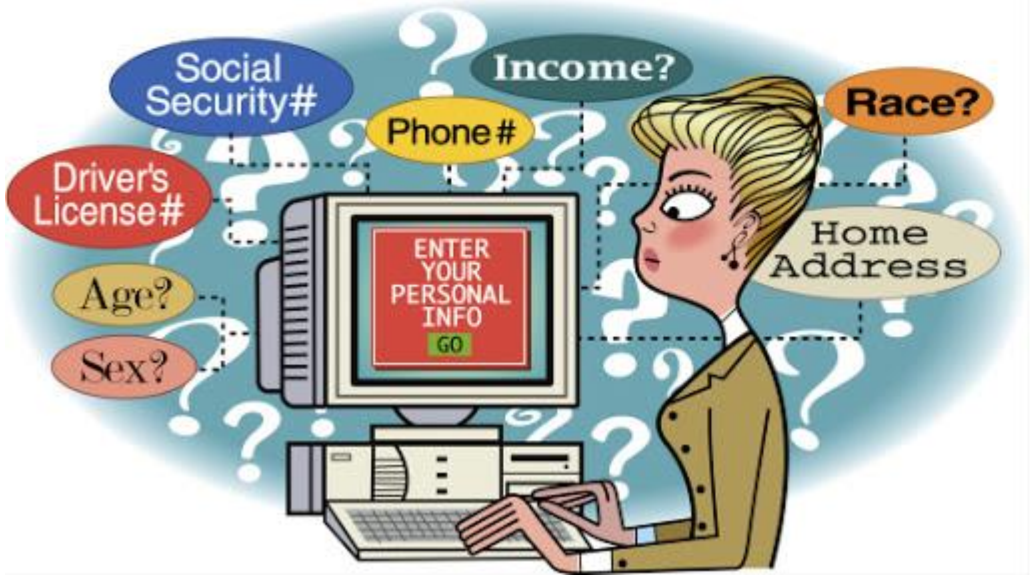


Transparency of what? (7. Usage)

7. Usage

Users may want to know what personal data a system is using, either to personalise outcomes or as data that can train the system to refine it or update it.
Knowing what personal data is used, they may then want to control that usage, perhaps to make their personalised results more accurate, or, more urgently, because they feel that usage violates their privacy, even though the data in question may already be a desired part of the system, such as a purchase or search history.

- There are grey areas here as well: collecting anonymised, highly detailed information about trips made by autonomous vehicles — how often the car brakes or swerves, for example — could be important to optimizing traffic for safety or fuel efficiency.
- Regulators may face some difficult decisions as well as drawing relatively obvious lines



Transparency by or for whom?

Because transparency has costs and risks, it matters **who** gets to see what is illuminated. When considering regulating transparency, the potential viewers include:

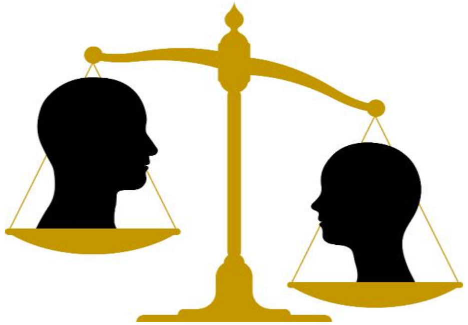
❖Everyone: fully open access to data, algorithms, outcomes, etc.

❖Regulatory authorities .

❖Third-party forensic analysts whose reports are made public, made selectively public, or kept private.

❖Researchers, possibly limited to those affiliated with accredited organisations and/or funding bodies.

Transparency - why?



We want systems to be transparent to help achieve important social goals related to accountability.

We want transparency of algorithmic system's **data and algorithms** in order to:

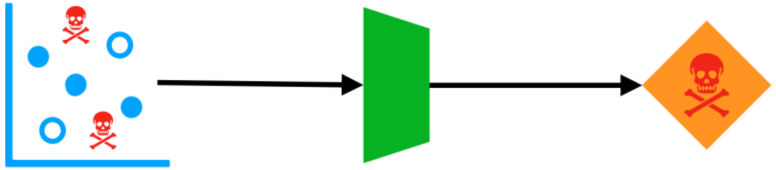
❖**Check for bias in the data and algorithms** that affects the fairness of the system.

❖Check that the system is drawing **inferences from relevant and representative data**.

❖See if we can learn anything from the machine's way of connecting and weighting the data

❖Look for, and **fix, bugs**.

❖**Guard against** malicious/adversarial **data injection**.



Poisoned Training Data

Poisoned Model

Transparency - why?

We want the **hierarchy of goals and outcomes** to be transparent so:

❖It can be **debated** and possibly regulated.

❖Regulators and the public can assess how well an algorithmic system has performed relative to its goals and compared to the pre-algorithmic systems it may be replacing or supplementing.

We want an **organisation's compliance status** to be public so:

❖Regulators can **hold the organisation accountable** in case of failure.

❖The public can evaluate the trustworthiness of the organisation, so people can make informed decisions as users about the services offered, and so citizens can become better informed about the benefits, risks, and trade-offs of algorithmic-based services overall.

Transparency by or for whom?

- Some of these potential costs can be mitigated by choosing where and how transparency interventions are necessary.
  - For example, rather than providing direct public access to the data being used to train a machine learning system, independent data scientists could examine the data in private and publish the conclusions of their forensic research.

- Transparency is not an absolute good and thus needs to be negotiated depending on its purpose and the balance of benefits and costs.
  - if an individual believes that s/he has been discriminated against by a black-box algorithmic system, but there is no evidence of systematic bias, the system might be tested to see if discriminatory factors were determinative in that particular outcome.
  - Such testing might not require transparency. For example, inputting counterfactual data — say, a loan application in which only a factor is changed at a time — can identify the impact of possibly prejudicial data without requiring full transparency.

Trade-offs Between Privacy And Transparency

- Search platforms tend to give little information about exactly what criteria their algorithms use.

- Therefore, algorithmic transparency can be at odds with the public (and commercial) interest in producing reliable, accurate search results.

- This can be addressed by keeping the algorithmic inspection limited to trusted experts who are not permitted to disclose what they learn. This of course also has some risks: disclosure by accident or corruption.

- Data Privacy could be compromised.
  - The data used to train a model is typically similar to that used by the model -- and in cases where this is data about individuals, the training data may be protected.
  - It may be possible to 'reverse engineer' a model to determine the data used to construct it, thus violating privacy.

Summary

- Transparency is a tool to be used responsibly, which means accepting that applying it means being sensitive to the complex contexts in which it is used, and the balance of benefits and harms its use inevitably entails. Transparency is a tool. As with any tool, whether and how it should best be used depends upon the context:
  - ❖The **goals** of requiring transparency.
  - ❖Which **elements** of the process should be made transparent.
  - ❖What **type of transparency** is most beneficial and least costly.
  - ❖**Alternative ways** of achieving the goal.
  - ❖A judgement of the potential trade-off between risk and the benefit an AI system could bring compared with the system it is replacing or augmenting.

Auditing Algorithms

- First, consider that algorithms can be manipulated in ways that do not disadvantage their users directly or obviously.
- Second, algorithmic manipulation may be societally problematic and deserving of scrutiny even if it is not illegal.
  - This is a significant observation because the majority of scholarship that has considered algorithmic discrimination in the past has done so from the perspective of law and regulation.
- **As algorithms (and computers) become more common in the implementation of all technological systems, studying the world means studying algorithms (auditing).**

Auditing Algorithms

- There are two modes of algorithmic audit:
  - **Direct auditing** consists of code audits and other more traditional efforts, which are effective on machine learning systems and models that human auditors can deconstruct and interpret.
  - **Indirect auditing** entails feeding sets of data that vary widely into an algorithm to test the outputs for signs of bias, anomalous behavior or other undesirable results.

Auditing Algorithms

- Algorithmic audits are new—so new that standardized methodologies don't yet exist. For now, researchers generally agree on a few basic guidelines for algorithmic audits.

- **Check your data: Reduce the bias**

  ❖Understanding what data has been collected and its relevance to the problem you're trying to solve, the integrity and accuracy of the data and the process undertaken to clean the raw data.

  ❖Having a clear understanding of what data may actually be a proxy for bias

- **Check how the algorithm works**

  ❖Understanding the components and weighting of an algorithm makes it significantly easier to perform an audit.

  ❖Asking what steps developers have taken to ensure the model is not biased against different protected groups.

Auditing Algorithms

- **Run your own “sniff test”**

  ❖Run some test data points through an algorithm and check the result. Running these simple “sniff tests” can often be useful in spotting the worst cases of bias quickly and affordably.

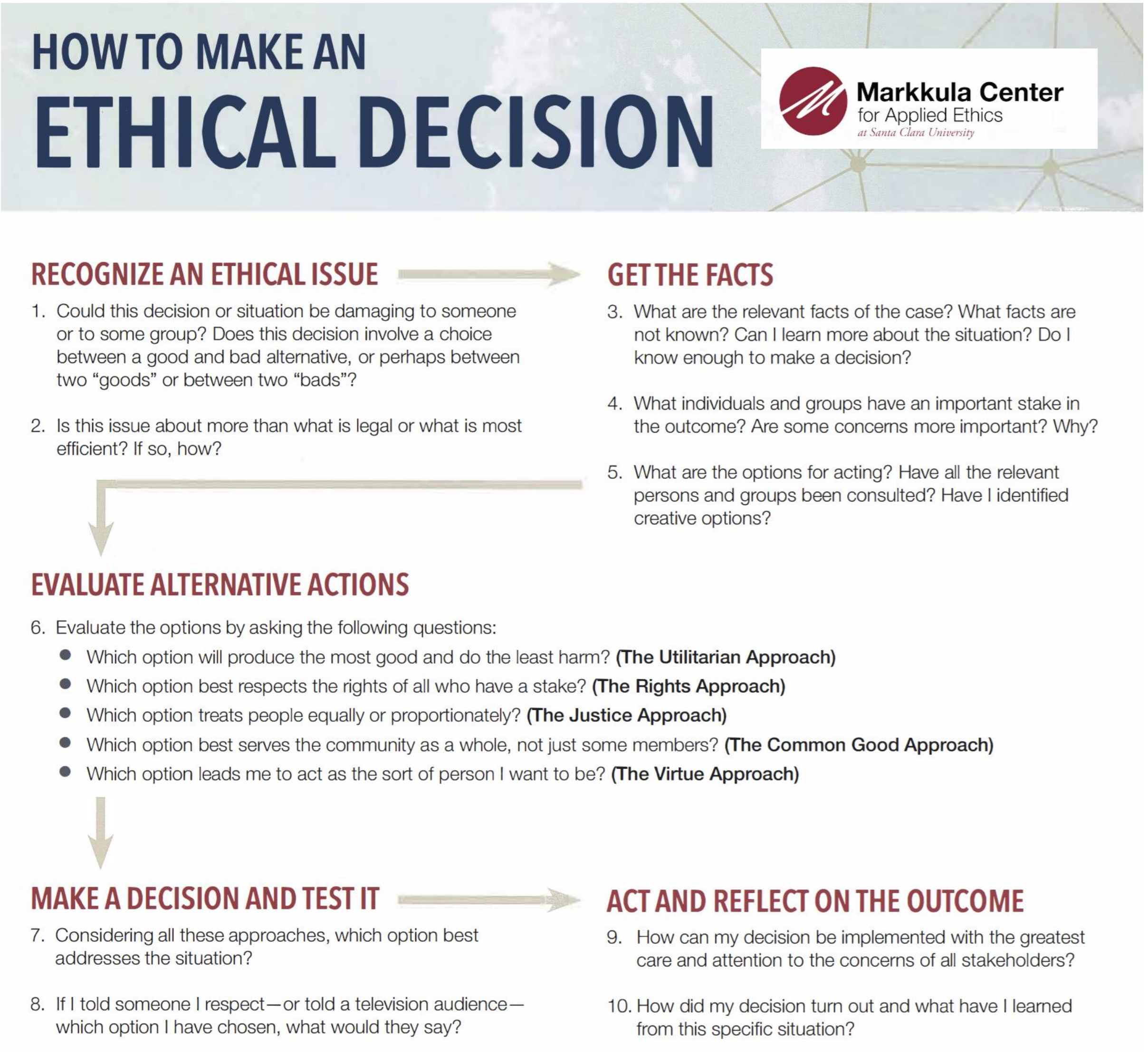
- **Request a full audit**

  ❖There are various levels of algorithmic audits. A company can hire an outside firm to come in and lead the effort. If the company has an internal data team that is developing data models, it can institute a basic auditing process that involves an extra layer of testing outcomes to look for signs of bias.

- **Audit regularly**

  ❖Bias—or just simply unfair results—may not be obvious initially and may only emerge as a model evolves or is trained on new data. Developers, who are used to writing code, checking it for errors, and then shipping it, aren’t used to having to work this way.

Frameworks for conducting ethical and legal algorithm audits



HOW TO MAKE AN ETHICAL DECISION

Markkula Center for Applied Ethics at Santa Clara University

RECOGNIZE AN ETHICAL ISSUE

1. Could this decision or situation be damaging to someone or to some group? Does this decision involve a choice between a good and bad alternative, or perhaps between two "goods" or between two "bads"?
2. Is this issue about more than what is legal or what is most efficient? If so, how?

GET THE FACTS

3. What are the relevant facts of the case? What facts are not known? Can I learn more about the situation? Do I know enough to make a decision?
4. What individuals and groups have an important stake in the outcome? Are some concerns more important? Why?
5. What are the options for acting? Have all the relevant persons and groups been consulted? Have I identified creative options?

EVALUATE ALTERNATIVE ACTIONS

6. Evaluate the options by asking the following questions:
   - Which option will produce the most good and do the least harm? **(The Utilitarian Approach)**
   - Which option best respects the rights of all who have a stake? **(The Rights Approach)**
   - Which option treats people equally or proportionately? **(The Justice Approach)**
   - Which option best serves the community as a whole, not just some members? **(The Common Good Approach)**
   - Which option leads me to act as the sort of person I want to be? **(The Virtue Approach)**

MAKE A DECISION AND TEST IT

7. Considering all these approaches, which option best addresses the situation?
8. If I told someone I respect—or told a television audience—which option I have chosen, what would they say?

ACT AND REFLECT ON THE OUTCOME

9. How can my decision be implemented with the greatest care and attention to the concerns of all stakeholders?
10. How did my decision turn out and what have I learned from this specific situation?

https://www.scu.edu/media/ethics-center/ethical-decision-making/A-Framework-for-Ethical-Decision-Making.pdf

Benefits and risks from algorithm audits

- Positive
  - find biases
  - keeping biases out of algorithms
  - How might the algorithm be abused?’
  - ‘What harm might it cause on a portion of society?’

- Negative
  - exposing your algorithms may destroy your competitive advantage.
  - for the people who are tech savvy, publishing the algorithm might allow them to understand the decision-making process in detail, thus allowing them effectively to game the system.

https://www.ft.com/content/879d96d6-93db-11e8-95f8-8640db9060a7

Tools for detecting Algorithmic Bias in AI

- Pymetrics: Audit AI
- DataScience.com Labs: Skater
- Google: What-If Tool
- IBM: AI Fairness 360 Open Source Toolkit
- Accenture: Teach & Test AI Framework

Q & A