# The lecture will begin shortly

Please, mute your microphones and turn-off your camera

# AI Ethics

AI for People workshop

# Intro

1. **My background**

# Intro

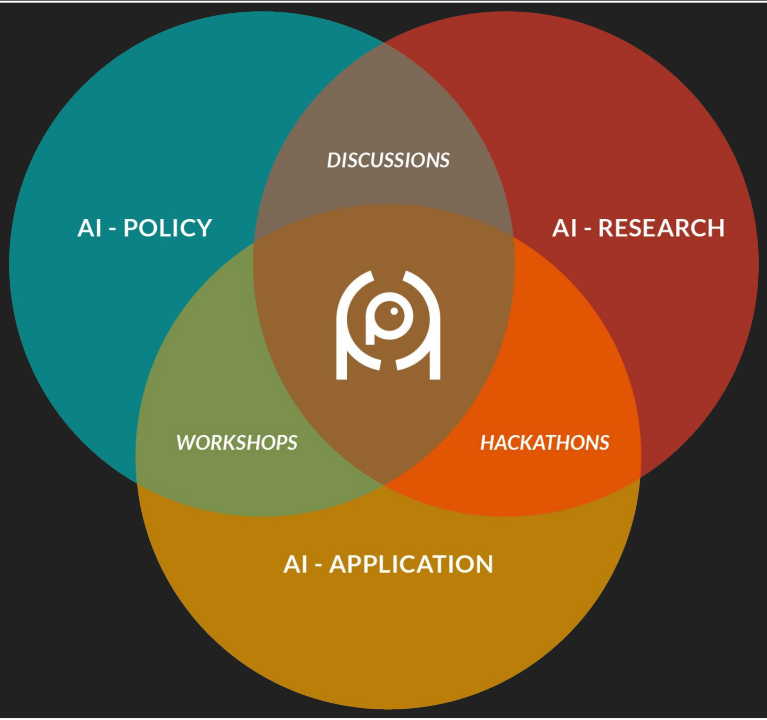
1. **My background**
   a. Philosophy & mathematics
   b. Public Policy & Philosophy
   c. AI Policy
   d. and….

# Intro

1. **My background**
   a. Philosophy & mathematics
   b. Public Policy & Philosophy
   c. AI Policy
   d. and…. AI for People!

# Intro

1. **My background**
   a. Philosophy & mathematics
   b. Public Policy & Philosophy
   c. AI Policy
   d. and…. AI for People!



**(AI → Society)**

**or**

**(Society → AI)?**

# Agenda

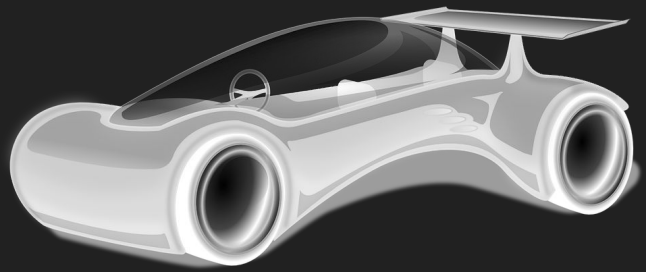
1. **Intro**
    a. AI
    b. Ethics
2. **Ethical Implications**
    a. problems → values → actions
3. **Bias, A hands-on case study**
    a. The COMPASS algorithm

# Agenda

1. Intro
    a. AI
    b. Ethics
2. Ethical Implications
    a. problems → values → actions
3. Bias, A hands-on case study
    a. The COMPASS algorithm

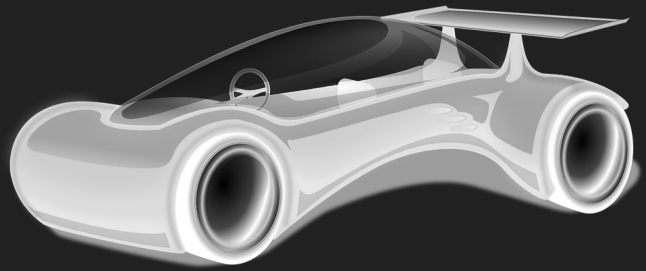
# AI

1. **What is AI and what is not?**
    a. **self-driving car vs**



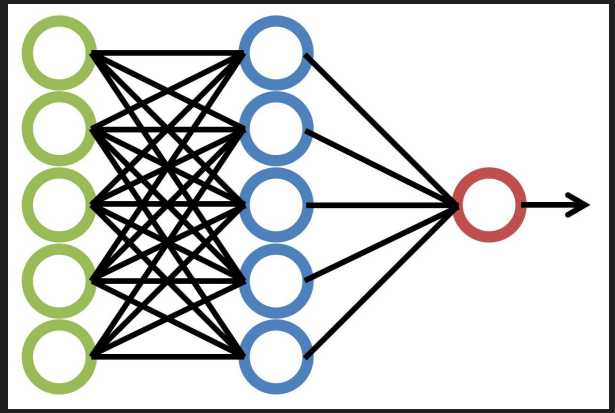
# AI

1. **What is AI and what is not?**
    a. **self-driving car vs facebook feed**



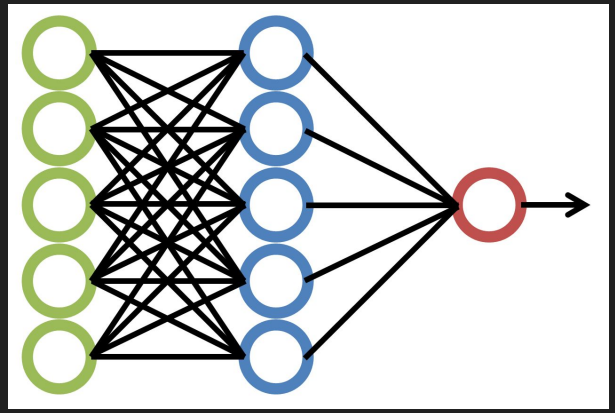
# AI

1. **What is AI and what is not?**
    a. **algorithm vs**



# AI

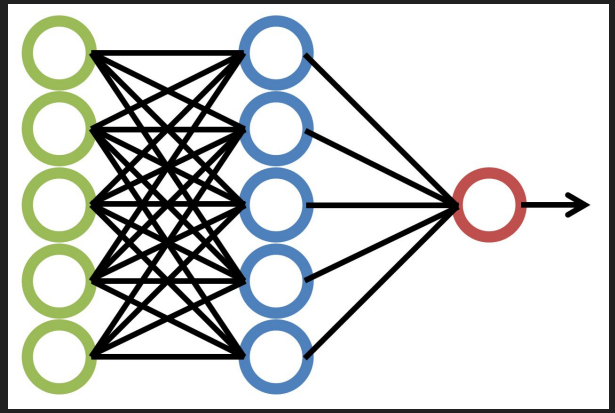
1. **What is AI and what is not?**
    a. **algorithm vs equation (?)**

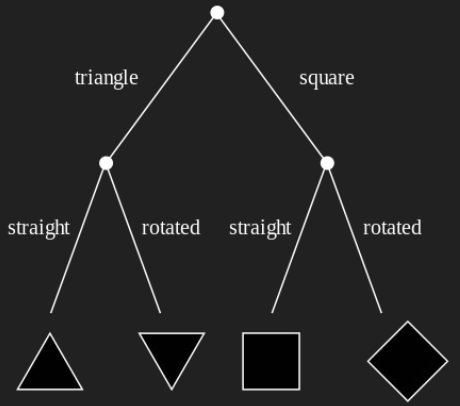


$$f(x) = ax^3 + bx^2 + cx + d$$

# AI

1. **What is AI and what is not?**
    a. **neural network vs decision tree**



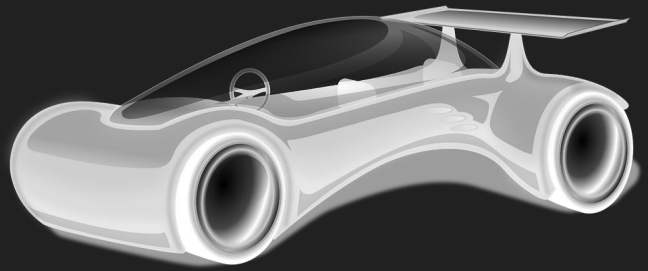


# AI

1. **There are:**
    a. **Components of AI**
    b. **Types of algorithms**
    c. **Types of tasks**
    d. **AI system lifecycle**

# AI

1. **There are:**
    a. **Components of AI (Hardware, Software...)**



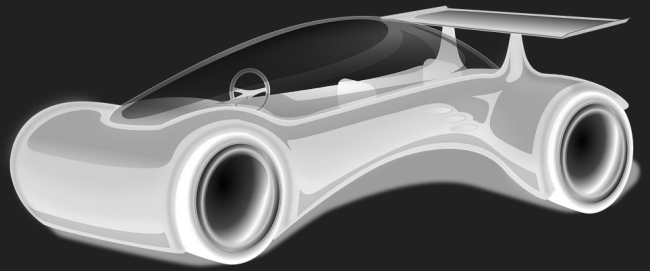
Hardware

Software

# AI

1. **There are:**
    a. **Components of AI (Hardware, Software… DATA!)**

DATA!

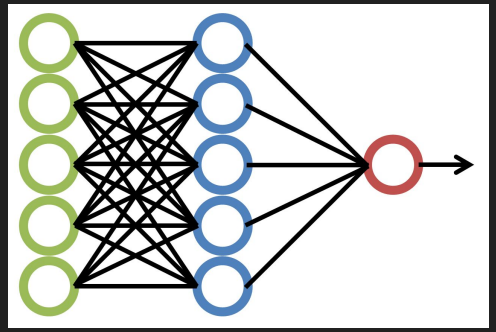


Hardware

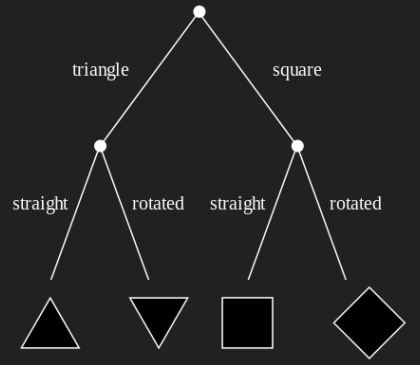
Software

# AI

1. **There are:**
   a. **Components of AI**
   b. **Types of algorithms**



Neural Networks



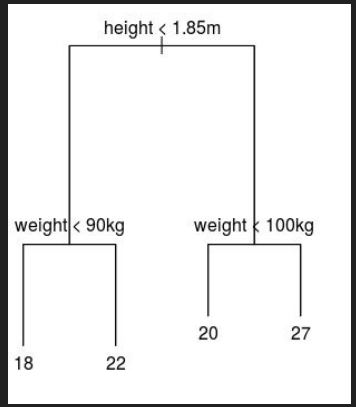

Decision Tree Classifier

# AI

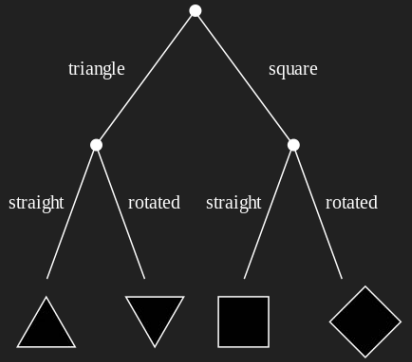
1. **There are:**
    a. **Types of AI (Software vs hardware)**
    b. **Types of algorithms (RNN, GANs..)**
    c. **Types of tasks**

# AI

1. **There are:**
    a. **Components of AI**
    b. **Types of algorithms**
    c. **Types of tasks**
        i. **Classification**
        ii. **Regression**
        iii. **Segmentation**
        iv. **Network Analysis**
        v. **…**




Regression



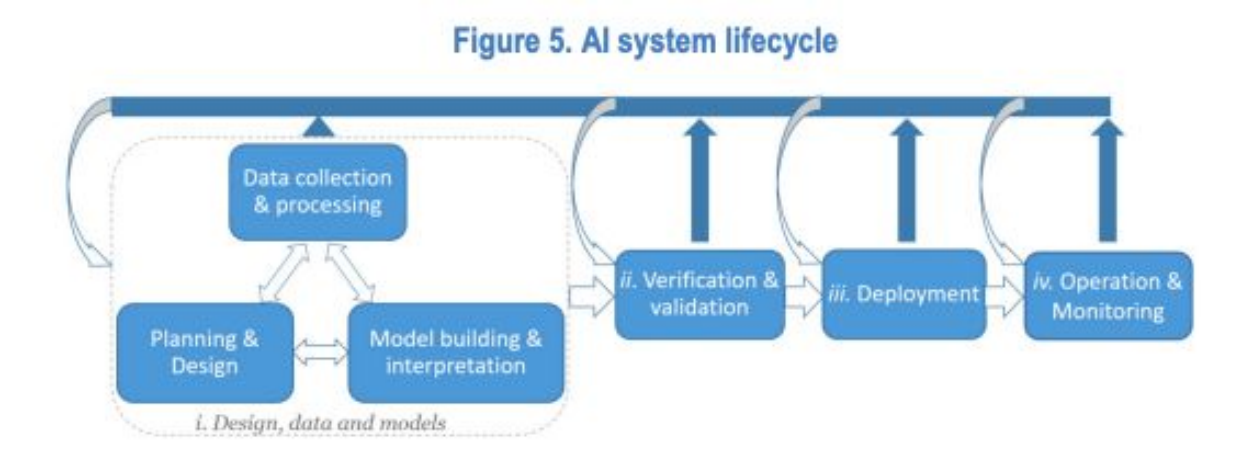

Classification

Regression image from: https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac

# AI

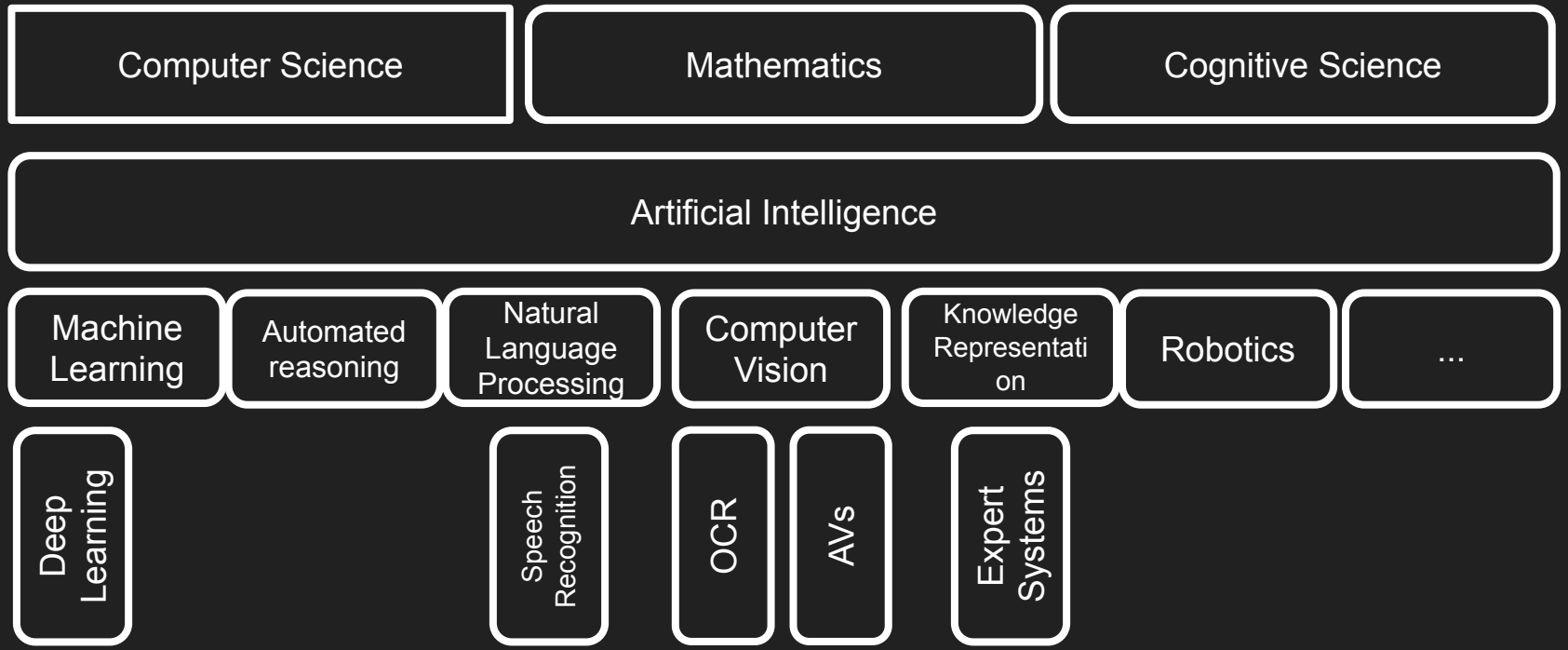
1. **There are:**
    a. **Components of AI**
    b. **Types of algorithms**
    c. **Types of tasks**
    d. **AI system lifecycle**

# AI




Source: OECD (2019), "Scoping the OECD AI principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)", *OECD Digital Economy Papers*, No. 291, OECD Publishing, Paris, https://doi.org/10.1787/d62f618a-en.

# What is AI? - AI Field




Source: Nicolas Moes ©

# Agenda

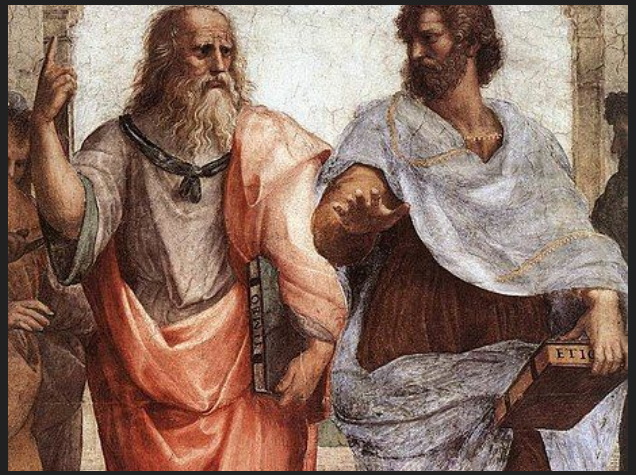
1. Intro
    a. AI
    b. Ethics
2. Ethical Implications
    a. problems → values → actions
3. Bias, A hands-on case study
    a. The COMPASS algorithm

# Ethics

1. **Virtue Ethics**
2. **Deontology**
3. **Consequentialism**
4. **Tensions between those?**

# Virtue Ethics

1. **Plato and Aristotle in the West, Mencius and Confucius in the East**



# Virtue Ethics

*arête* (excellence or virtue)

# Virtue Ethics

*arête* (excellence or virtue) + *phronesis* (practical or moral wisdom) =

# Virtue Ethics

*eudaimonia* (usually translated as happiness or flourishing)

# Virtue Ethics

*arête + phronesis = eudaimonia*

# Deontology

duty (*deon*) + science (or study) of (*logos*)

1. **Kant**

# Deontology

duty (*deon*) + science (or study) of (*logos*)

1. **Kant**
2. Ten Commandments & Universal Declaration of Human Rights

# Deontology

duty (*deon*) + science (or study) of (*logos*)

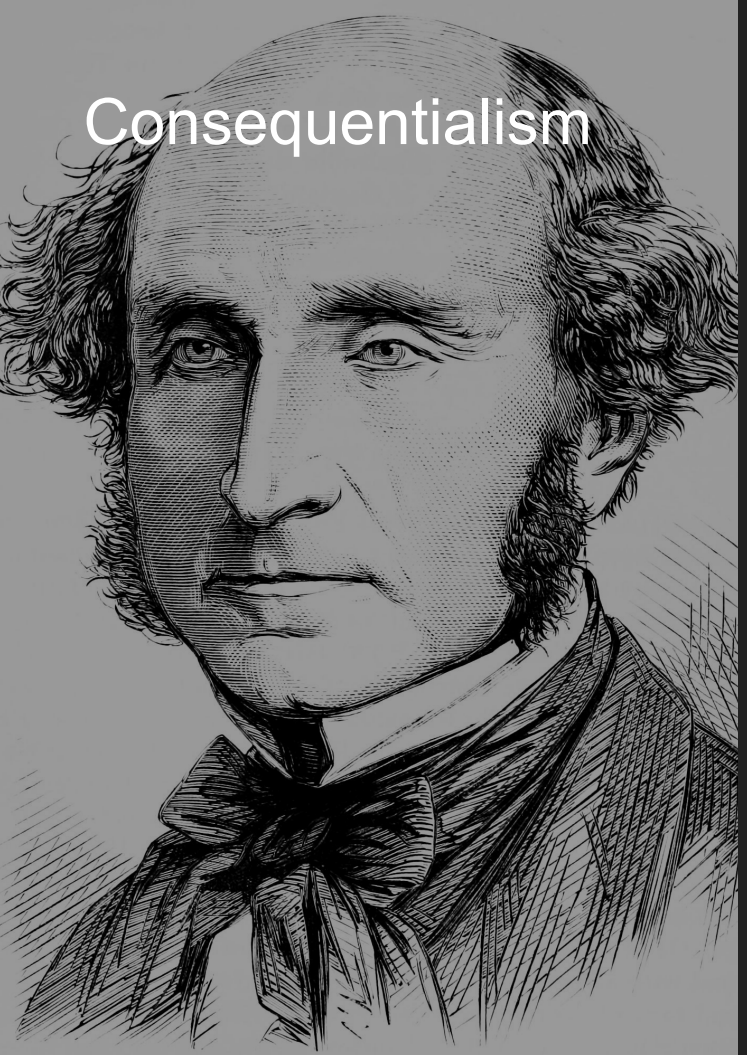
1. **Kant**
2. *The categorical imperative*

# Deontology

*The categorical imperative*

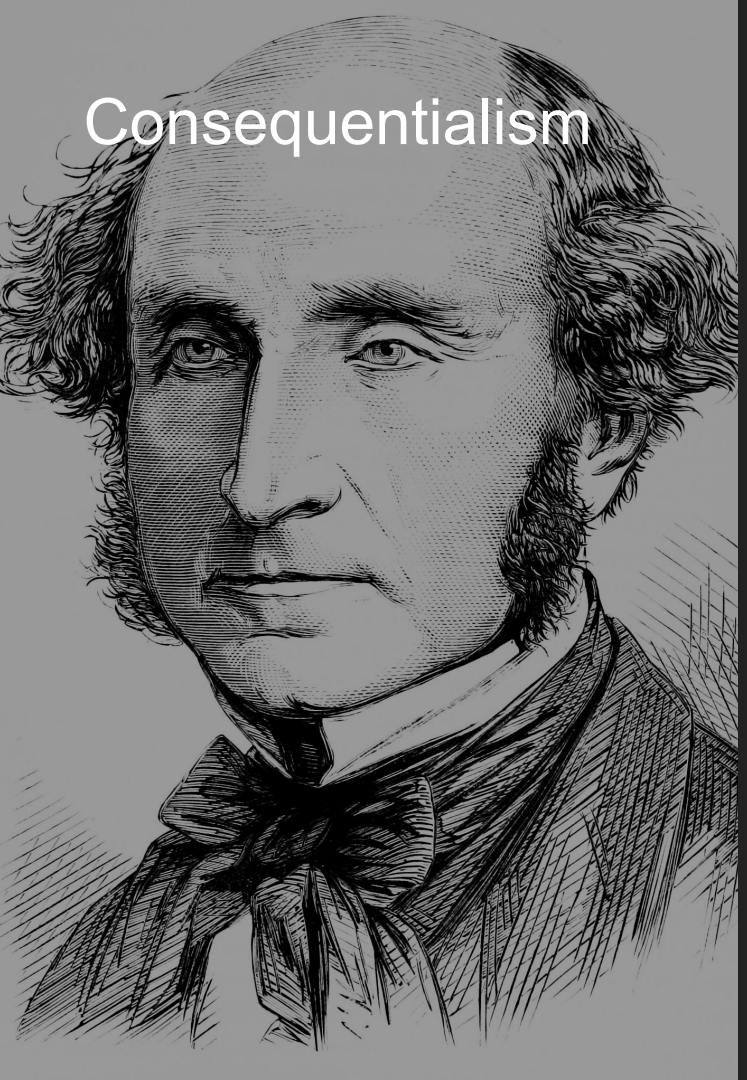
But Consequentialists...

# Consequentialism



1. **i.e. Utilitarianism, The maximization of happiness**
2. **Mill, Bentham & Sidwick**

# Consequentialism



1. **i.e. Utilitarianism, The maximization of happiness**
2. **Mill, Bentham & Sidwick**

Consequentialism → Hedonistic Act

# Consequentialism

Hedonistic Act $\longrightarrow$ Max (Happiness)

# Ethics

1. **Virtue Ethics**
2. **Deontology**
3. **Consequentialism**
4. **Tensions between those?**
    a. **Which strand would support the thought ‘How much evil must we do before doing good?’**
        i. **Virtue ethics →**
        ii. **Deontology →**
        iii. **Utilitarianism →**

# Ethics

1. **Virtue Ethics**
2. **Deontology**
3. **Consequentialism**
4. **Tensions between those?**
    a. **Which strand would support the thought ‘How much evil must we do before doing good?’**
        i. **Virtue ethics → X**
        ii. **Deontology →**
        iii. **Utilitarianism →**

# Ethics

1. **Virtue Ethics**
2. **Deontology**
3. **Consequentialism**
4. **Tensions between those?**
    a. **Which strand would support the thought ‘How much evil must we do before doing good?’**
        i. **Virtue ethics → X**
        ii. **Deontology → X**
        iii. **Utilitarianism →**

# Ethics

1. **Virtue Ethics**
2. **Deontology**
3. **Consequentialism**
4. **Tensions between those?**
    a. **Which strand would support the thought ‘How much evil must we do before doing good?’**
        i. **Virtue ethics → X**
        ii. **Deontology → X**
        iii. **Utilitarianism → V**

# Ethics & AI?

1. **Can AI be ethical? (not our business today)**
2. **Can humans engage in AI ethically? (our business today)**

# Agenda

1. Intro
    a. AI
    b. Ethics
2. Ethical Implications
    a. problems → values → actions
3. Bias, A hands-on case study
    a. The COMPASS algorithm

# Ethical implications

# Ethical implications

**Algorithms**

**turn data into evidence for a given outcome**

# Ethical implications

**Algorithms**

**turn data into evidence for a given outcome**

**outcome used to trigger + motivate an action**

# Ethical implications

**Algorithms**

turn data into evidence for a given **outcome**

**outcome** used to trigger + motivate an **action**

**action** may not be ethically neutral

# Ethical implications

**Algorithms**

turn data into evidence for a given **outcome**

**outcome** used to trigger + motivate an **action**

**action** may not be ethically neutral

**AND**

works in **complex** & (semi)**autonomous** way

# Ethical Implications

**This generates…**

*problems → values → actions*

# Ethical Implications

**problems → values → actions**

1. Ethical AI/ML
2. Problems and challenges
3. The most ethical approach?
4. The lawful approach?

# Ethical Implications

**problems → values → actions**

1. Ethical AI/ML
2. Problems and challenges
3. The most ethical approach?
4. The lawful approach?

# Ethical AI/ML

*The debate about the ethical implications of AI dates from the 1960s (Samuel, Wiener, Turing)*

*Shift from symbolic systems to (Deep) NN and ML*

# Ethical AI/ML

*ML algorithms are powerful socio-technical constructs (Ananny & Crawford, 2018)*

*They raise concerns about people as much as about code*

*Ethical, robust and lawful (EC, ethical guidelines)*

# Ethical Implications

**problems** → **values** → **actions**

1. Ethical AI/ML
2. Problems and challenges
3. The most ethical approach?
4. The lawful approach?

# Ethical implications

**Algorithms**

**turn data into evidence for a given outcome**

**outcome used to trigger + motivate an action**

# Ethical implications

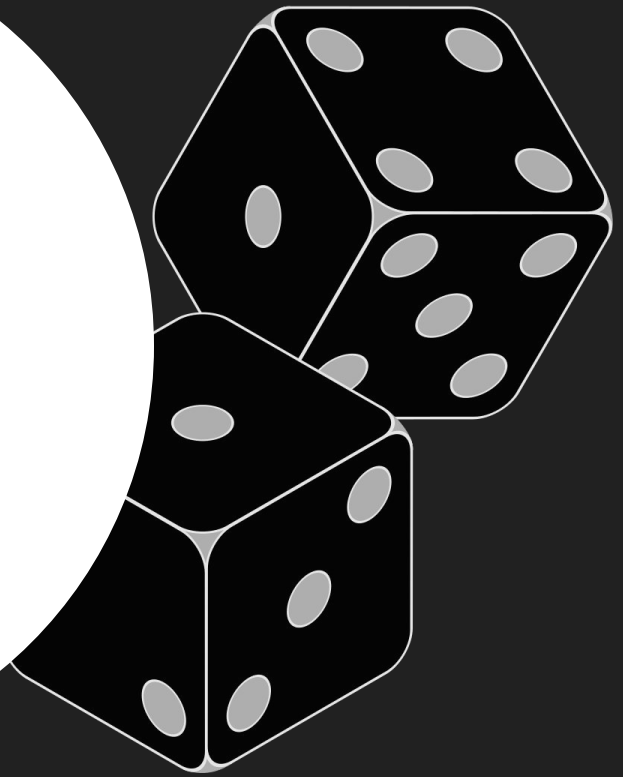
**Algorithms**

**turn data into evidence for a given outcome**

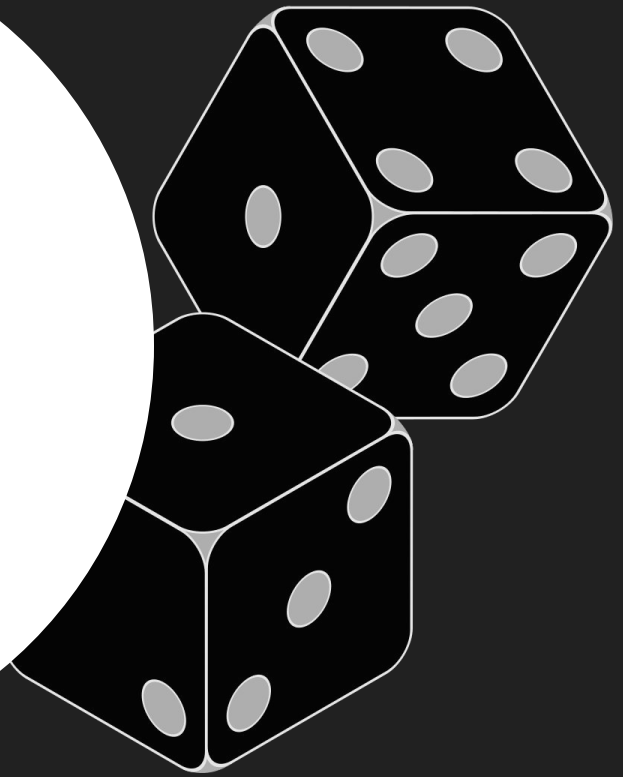
**outcome used to trigger + motivate an action**

**Inconclusive evidence**

# Inconclusive Evidence



Example:



# Ethical implications

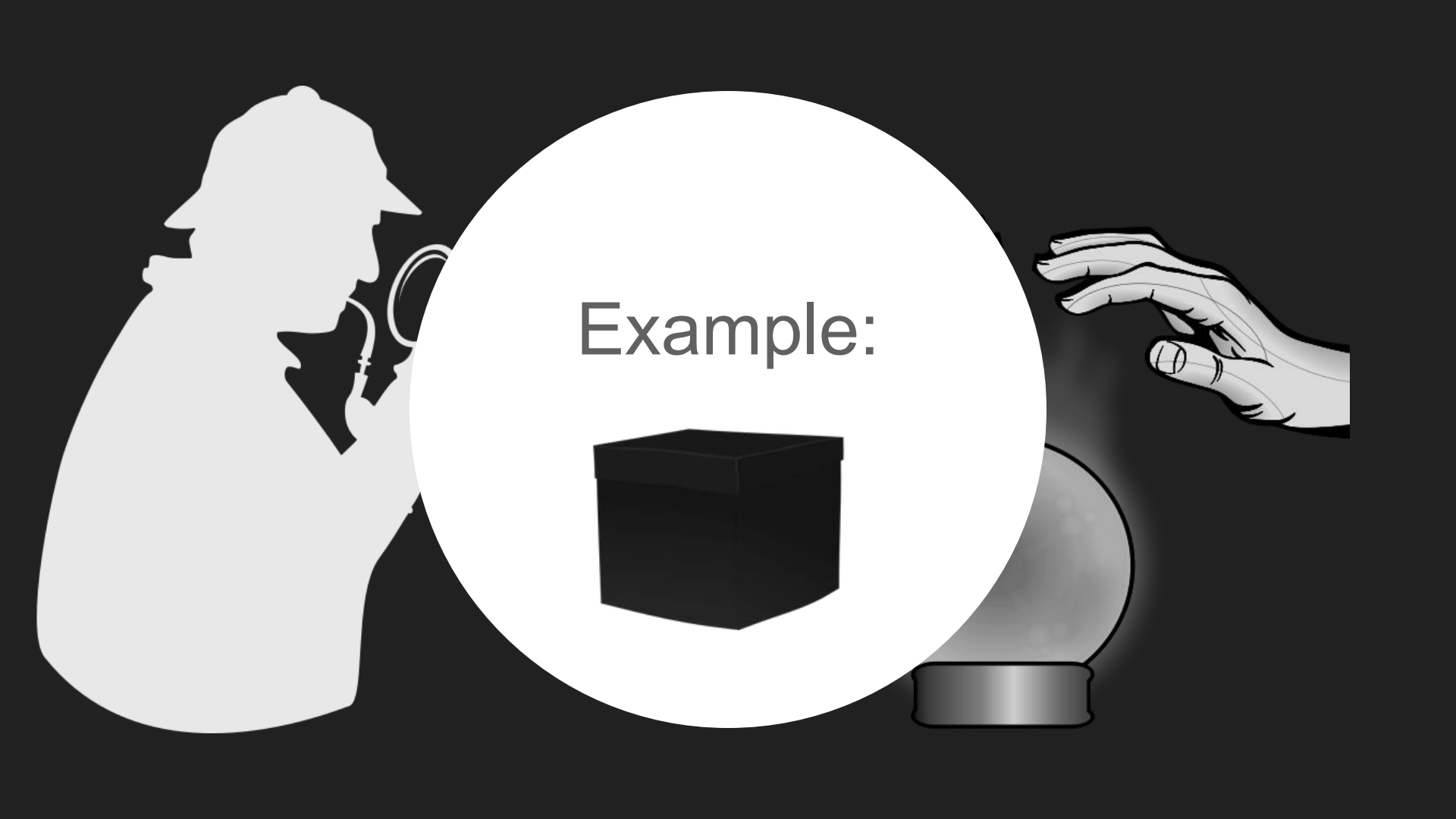
**Algorithms**

**turn data into evidence for a given outcome**

**outcome used to trigger + motivate an action**

**Inscrutable evidence**


Inscrutable
Evidence


Example:


Example:

# Ethical implications

**Algorithms**

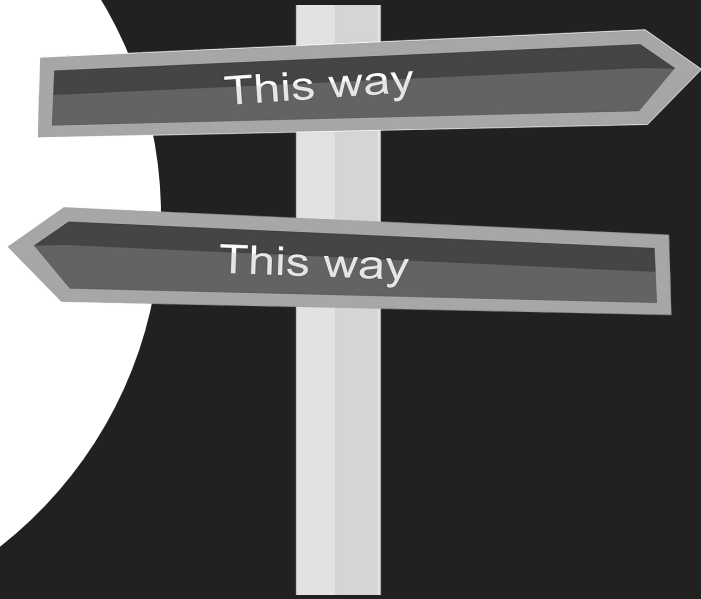
**turn data into evidence for a given outcome**

**outcome used to trigger + motivate an action**

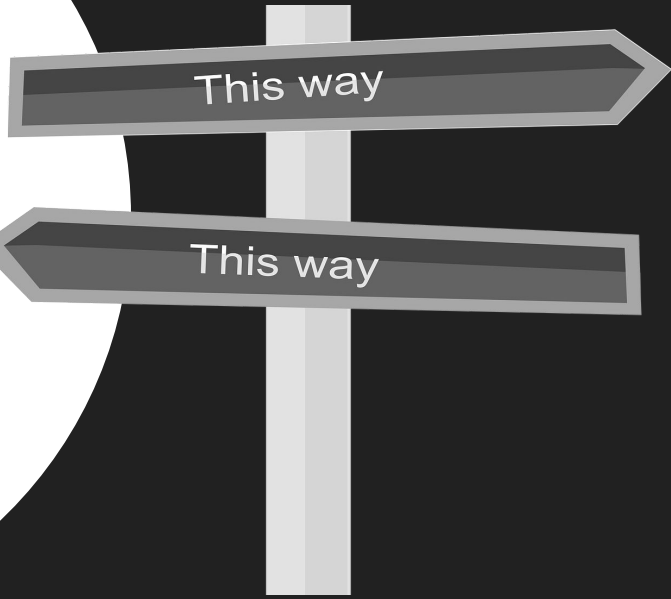
**action may not be ethically neutral**

**Misguided Evidence**

# Misguided Evidence



Example:



# Ethical implications

**Algorithms**

turn data into evidence for a given **outcome**

**outcome** used to trigger + motivate an **action**

**action** may not be ethically neutral

# Ethical implications

**Algorithms**

**turn data into evidence for a given outcome**

**outcome used to trigger + motivate an action**

**action may not be ethically neutral**

**Unfair Outcomes**


Unfair
Outcomes

Example:

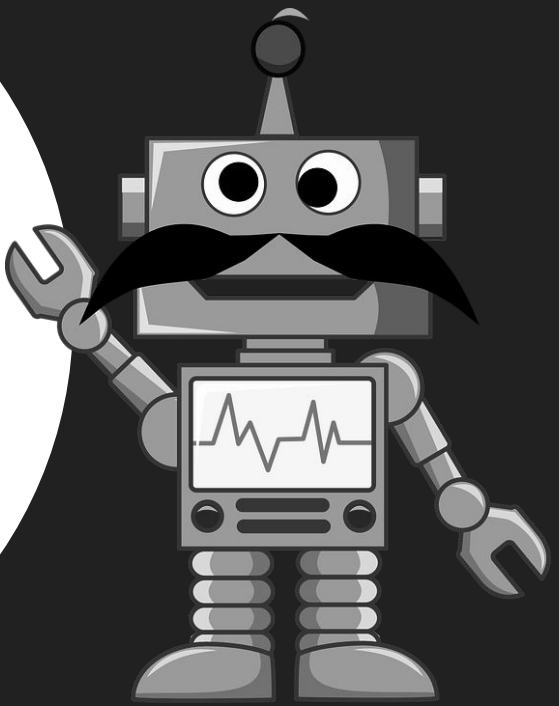
# Ethical implications

**Algorithms**

**work in complex & (semi)autonomous way**

**Transformative effects**

# Transformative Effects




Example:

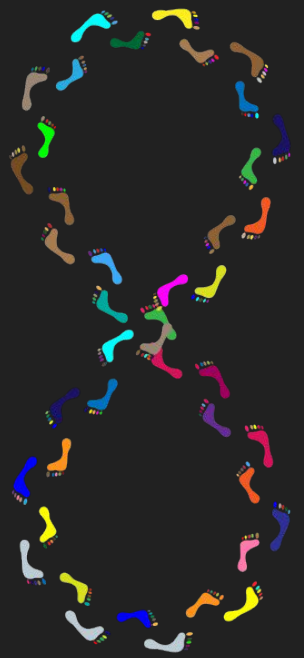
# Ethical implications

**Algorithms**

**work in complex & (semi)autonomous way**

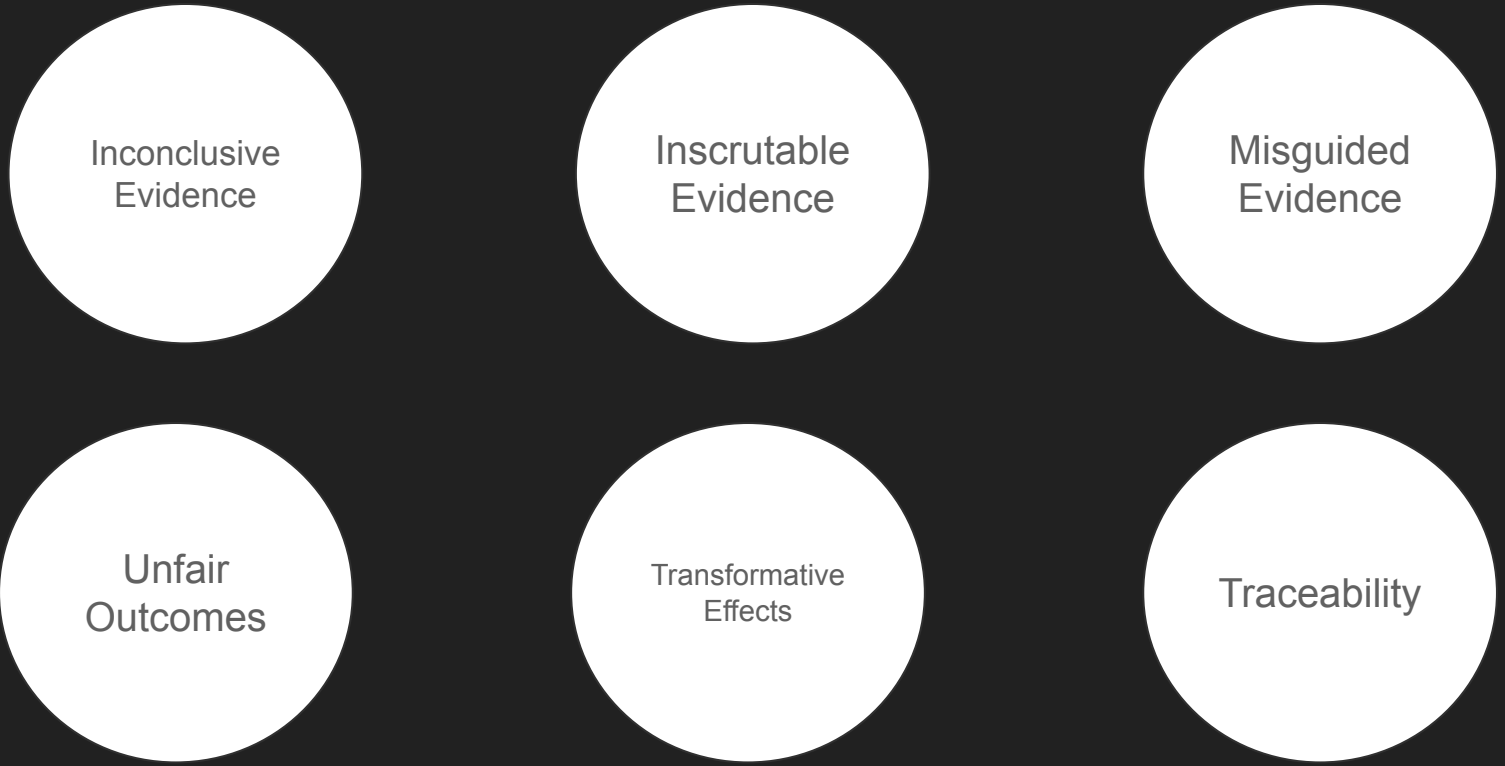
**Traceability**

Traceability



Example

# Ethical Problems



# Ethical Implications

**problems → values → actions**

1. Ethical AI/ML
2. Problems and challenges
3. The most ethical approach?
4. The lawful approach?

# The most ethical approach?

1. **The most ethical approach?**
    a. The “What”, a virtue ethics approach
        i. the values: Fairness, Bias, Explainability…

# The most ethical approach?

1. **The most ethical approach?**
    a. The “What”, a virtue ethics approach
        i. the values: Fairness, Bias, Explainability…
    b. The “How”, a pragmatic ethical approach
        i. The tools: tools for fairness, bias, explainability…

# The most ethical approach?

The ‘What’, a virtue ethics approach

1. *The principles put forward can, as abstractions, act as normative constraints on the do’s and don’ts of algorithmic use in society’* (Turilli, 2007)

# The most ethical approach?

The ‘What’, a virtue ethics approach

> *More than 70 documents (as ethical guidelines) in the last three years*

# The most ethical approach?

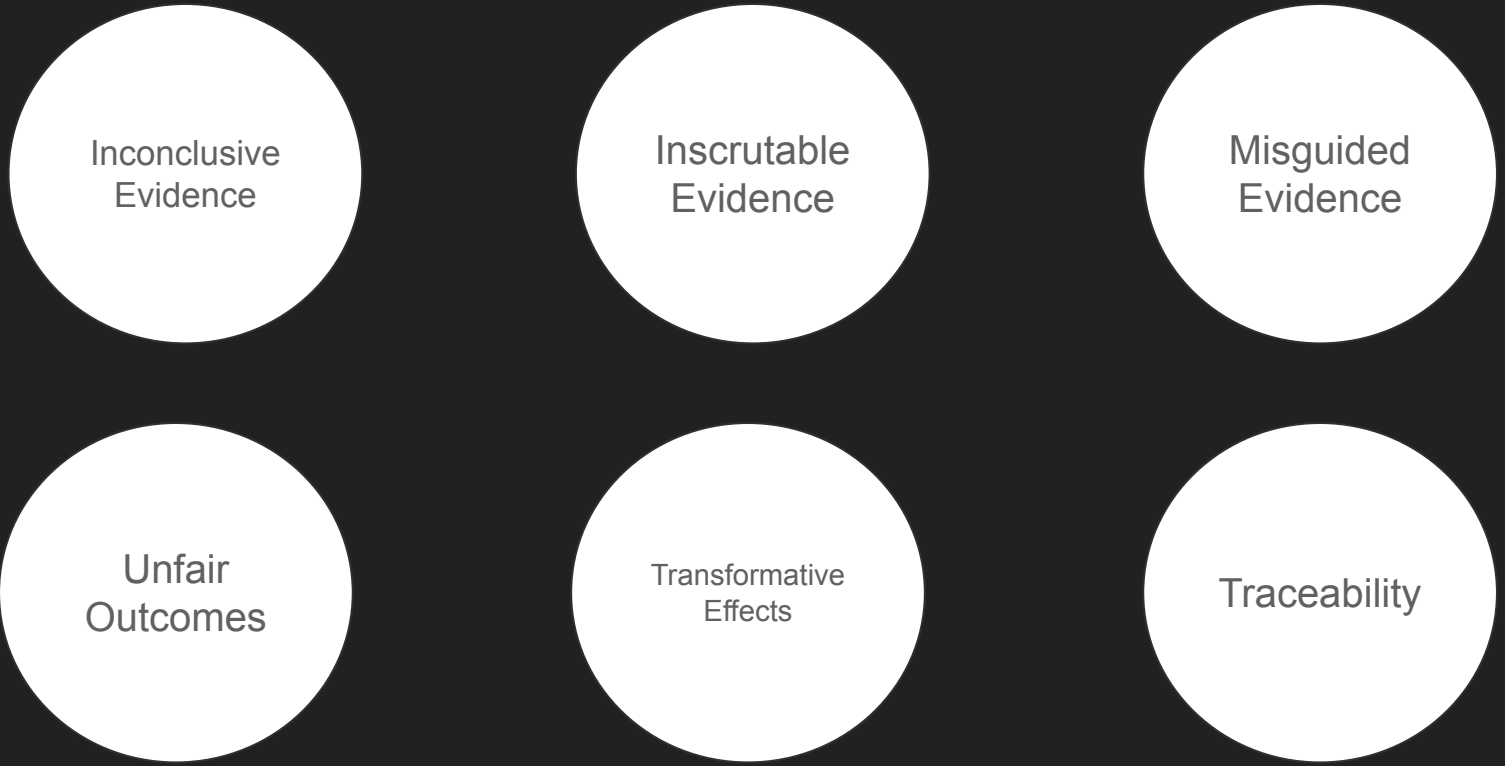
The ‘What’, a virtue ethics approach

*More than 70 documents (as ethical guidelines) in the last three years*

a. Industry (Google, IBM..), government (Montreal Declaration, HLEG from EC..), intergovernmental institutions (OECD), academia (Future of Life Institute, IEEE..)

# The “What”, the values

# Ethical Problems



# The “What”, the values

Inconclusive Evidence

# The “What”, the values

Inconclusive Evidence

Accuracy

# The “What”, the values

Inscrutable Evidence

# The “What”, the values

Inscrutable Evidence

↓

Explainability & Transparency

# The “What”, the values

Misguided Evidence

# The “What”, the values

Misguided Evidence

Bias

# The “What”, the values

Unfair Outcomes

# The “What”, the values

Unfair Outcomes

↓

Fairness

# The “What”, the values

Transformative Effects

# The “What”, the values

Transformative Effects

Privacy

# The “What”, the values

Traceability

# The “What”, the values

Traceability

Accountability

# The most ethical approach?

The ‘What’, a virtue ethics approach

> *More than 70 documents (as ethical guidelines) in the last three years*

# The most ethical approach?

The ‘What’, a virtue ethics approach

*More than 70 documents (as ethical guidelines) in the last three years*

Emerging Consensus!

# The most ethical approach?

The ‘What’, a virtue ethics approach

*More than 70 documents (as ethical guidelines) in the last three years*

Not a “How” for our “Whats”

# The most ethical approach?

*A review of 84 ethical AI documents by Jobin et al. (2019) found that no single principle featured in all of them*

# The most ethical approach?

*Nevertheless*

# The most ethical approach?

*Themes of transparency, justice and fairness, non-maleficence, responsibility and privacy appeared in over half*

| Ethical principle | Number of documents | Included codes |
|---|---|---|
| Transparency | 73/84 | Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing |
| Justice and fairness | 68/84 | Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution |
| Non-maleficence | 60/84 | Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion |
| Responsibility | 60/84 | Responsibility, accountability, liability, acting with integrity |
| Privacy | 47/84 | Privacy, personal or private information |
| Beneficence | 41/84 | Benefits, beneficence, well-being, peace, social good, common good |
| Freedom and autonomy | 34/84 | Freedom, autonomy, consent, choice, self-determination, liberty, empowerment |
| Trust | 28/84 | Trust |
| Sustainability | 14/84 | Sustainability, environment (nature), energy, resources (energy) |
| Dignity | 13/84 | Dignity |
| Solidarity | 6/84 | Solidarity, social security, cohesion |

Source: Royakkers, L., Timmer, J., Kool, L., & van Est, R. (2018). Societal and ethical issues of digitization. Ethics and Information Technology, 20(2), 127–142. https://doi.org/10.1007/s10676-018-9452-x

# The most ethical approach?

*Themes of privacy, security, autonomy, justice, human dignity, control of technology and the balance of powers, were recurrent (Royakkers et al., 2018)*

Source: Royakkers, L., Timmer, J., Kool, L., & van Est, R. (2018). Societal and ethical issues of digitization. Ethics and Information Technology, 20(2), 127–142. https://doi.org/10.1007/s10676-018-9452-x

| Ethical principle | Number of documents | Included codes |
|---|---|---|
| Transparency | 73/84 | Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing |
| Justice and fairness | 68/84 | Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution |
| Non-maleficence | 60/84 | Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion |
| Responsibility | 60/84 | Responsibility, accountability, liability, acting with integrity |
| Privacy | 47/84 | Privacy, personal or private information |
| Beneficence | 41/84 | Benefits, beneficence, well-being, peace, social good, common good |
| Freedom and autonomy | 34/84 | Freedom, autonomy, consent, choice, self-determination, liberty, empowerment |
| Trust | 28/84 | Trust |
| Sustainability | 14/84 | Sustainability, environment (nature), energy, resources (energy) |
| Dignity | 13/84 | Dignity |
| Solidarity | 6/84 | Solidarity, social security, cohesion |

Source: Royakkers, L., Timmer, J., Kool, L., & van Est, R. (2018). Societal and ethical issues of digitization. Ethics and Information Technology, 20(2), 127–142. https://doi.org/10.1007/s10676-018-9452-x

# The ‘How’, a pragmatic approach

Not a “How” for our “Whats”

# The ‘How’, a pragmatic approach

Not a “How” for our “Whats”

*‘to know not only what to do or not to do, but also how to do it, or avoid doing it’* (Alshammari & Simpson, 2017).

Source: Alshammari, M., & Simpson, A. (2017). Towards a principled approach for engineering privacy by design. In E. Schweighofer, H. Leitold, A. Mitrakas, & K. Rannenberg (Eds.), *Privacy technologies and policy* (Vol. 10518, pp. 161–177). Cham: Springer. https://doi.org/10.1007/978-3-319-67280-9_9.

# The ‘How’, a pragmatic approach

Not a “How” for our “Whats”

*‘The guidelines often suggest that technical solutions exist, but very few provide technical explanations’*

Source: Alshammari, M., & Simpson, A. (2017). Towards a principled approach for engineering privacy by design. In E. Schweighofer, H. Leitold, A. Mitrakas, & K. Rannenberg (Eds.), *Privacy technologies and policy* (Vol. 10518, pp. 161–177). Cham: Springer. https://doi.org/10.1007/978-3-319-67280-9_9.

# The ‘How’, a pragmatic approach

Not a “How” for our “Whats”

*79% of tech workers report that they would like practical resources to help them with ethical considerations (Miller & Coldicott, 2019)*

Source: Miller, C., & Coldicott, R. (2019). *People, power and technology: The tech workers’ view*. Retrieved from Doteveryone website: https://doteveryone.org.uk/report/workersview/.

# The ‘How’, a pragmatic approach

Not a “How” for our “Whats”

*Mapping needed to have a ‘how’ for every ‘what’*

Source: Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 1-28.

# The ‘How’, a pragmatic approach

https://www.aiforpeople.org/policies/

Source: Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 1-28.

# The ‘How’, a pragmatic approach

https://www.aiforpeople.org/policies/

Based on:

Table:
https://docs.google.com/document/d/1h6nK9K7qspG74_HyVlT0Lx97URM0dRoGbJ3ivPxMhaE/edit

Interactive tool:
https://projects.invisionapp.com/share/EDV8PZWN7WZ#/screens

Source: Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 1-28.

# Ethical Implications

**problems → values → actions**

1. Ethical AI/ML
2. Problems and challenges
3. The most ethical approach?
4. The lawful approach?

# The lawful approach?

1. Examples
2. Challenges
3. How to take part?

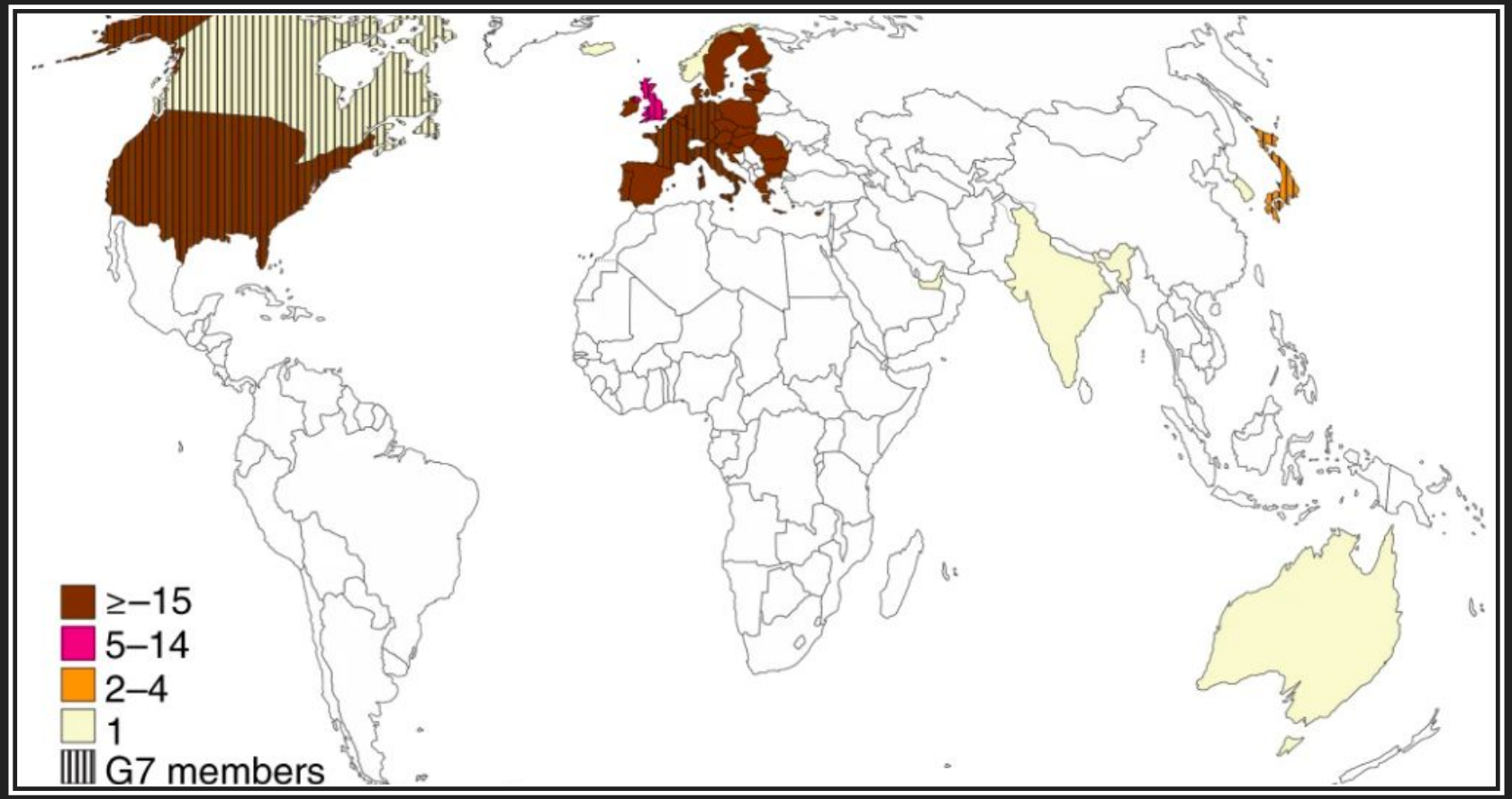
# The lawful approach?

1. Examples
2. Challenges
3. How to take part?

# The lawful approach?

*More than 70 documents (as ethical guidelines) in the last three years*

*Government (Montreal Declaration, HLEG from EC, Germany, Italy...), intergovernmental institutions (OECD), academia (Future of Life Institute, IEEE..), industry (Google, IBM..)*




Source: Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

# The lawful approach?

*More than 70 documents (as ethical guidelines) in the last three years*

*Government (Montreal Declaration, HLEG from EC, Germany, Italy...), intergovernmental institutions (OECD), academia (Future of Life Institute, IEEE..), Industry (Google, IBM..)*

# The lawf

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment


*nical guidelines) in*
*ars*

*ation, HLEG from*
*governmental*
*a (Future of Life*
*Google, IBM..)*

# The la

## White Paper on Artificial Intelligence
## A European approach to excellence and trust

Artificial Intelligence is developing fast. It will change our lives by improving healthcare (e.g. making diagnosis more precise, enabling better prevention of diseases), increasing the efficiency of farming, contributing to climate change mitigation and adaptation, improving the efficiency of production systems through predictive maintenance, increasing the security of Europeans, and in many other ways that we can only begin to imagine. At the same time, Artificial Intelligence (AI) entails a number of potential risks, such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes.

Against a background of fierce global competition, a solid European approach is needed, building on the European strategy for AI presented in April 2018[1]. To address the opportunities and challenges of AI, the EU must act as one and define its own way, based on European values, to promote the development and deployment of AI.

The Commission is committed to enabling scientific breakthrough, to preserving the EU's technological leadership and to ensuring that new technologies are at the service of all Europeans – improving their lives while respecting their rights.

Commission President Ursula von der Leyen announced in her political Guidelines[2] a coordinated European approach on the human and ethical implications of AI as well as a reflection on the better use of big data for innovation.

Thus, the Commission supports a regulatory and investment oriented approach with the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology. The purpose of this White Paper is to set out policy options on how to achieve these objectives. It does not address +the development and use of AI for military purposes.The Commission invites Member States, other European institutions, and all stakeholders, including industry, social partners, civil society organisations, researchers, the public in general and any interested party, to react to the options below and to contribute to the Commission's future decision-making in this domain.

*idelines) in*

*LEG from*
*mental*
*e of Life*
*IBM..)*

# The lawful approach?

*More than 70 documents (as ethical guidelines) in the last three years*

*Government (Montreal Declaration, HLEG from EC, Germany, Italy...), intergovernmental institutions (OECD), academia (Future of Life Institute, IEEE..), Industry (Google, IBM..)*

## What are the OECD Principles on AI?

The OECD Principles on Artificial Intelligence promote artificial intelligence (AI) that is innovative and trustworthy and that respects human rights and democratic values. They were adopted in May 2019 by OECD member countries when they approved the **OECD Council Recommendation on Artificial Intelligence**. The OECD AI Principles are the first such principles signed up to by governments. Beyond OECD members, other countries including Argentina, Brazil, Costa Rica, Malta, Peru, Romania and Ukraine have already adhered to the AI Principles, with further adherents welcomed.

The OECD AI Principles set standards for AI that are practical and flexible enough to stand the test of time in a rapidly evolving field. They complement existing OECD standards in areas such as privacy, digital security risk management and responsible business conduct.

In June 2019, the **G20 adopted human-centred AI Principles** that draw from the OECD AI Principles.

## The OECD AI Principles

The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being.
- AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.
- There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.
- AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
- Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles.

# The lawful approach?

*More than 70 documents (as ethical guidelines) in the last three years*

*Government (Montreal Declaration, HLEG from EC, Germany, Italy...), intergovernmental institutions (OECD), academia (Future of Life Institute, IEEE..), Industry (Google, IBM..)*

## Ethics and Values

6) **Safety:** AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

7) **Failure Transparency:** If an AI system causes harm, it should be possible to ascertain why.

8) **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

9) **Responsibility:** Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

10) **Value Alignment:** Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.

11) **Human Values:** AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

12) **Personal Privacy:** People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.

13) **Liberty and Privacy:** The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.

14) **Shared Benefit:** AI technologies should benefit and empower as many people as possible.

# The lawful approach?

*More than 70 documents (as ethical guidelines) in the last three years*

*Government (Montreal Declaration, HLEG from EC, Germany, Italy...), intergovernmental institutions (OECD), academia (Future of Life Institute, IEEE..), Industry (Google, IBM..)*

# General recommended practices for AI

Reliable, effective user-centered AI systems should be designed following general best practices for software systems, together with practices that address considerations unique to machine learning. Our top recommendations are outlined below, with additional resources for further reading.



Back to top ↑

## Recommended practices

- Use a human-centered design approach
- Identify multiple metrics to assess training and monitoring
- When possible, directly examine your raw data

# The lawful approach?

1. Examples
2. Challenges
3. How to take part?

# The lawful approach?

Problems:

1. None is binding

# The lawful approach?

Problems:

1. None is binding → GDPR?

# The lawful approach?

Problems:

1. None is binding → GDPR?
2. Quantity over quality?

# The lawful approach?

Problems:

1. None is binding → GDPR?
2. Quantity over quality?
3. Too Western-centric?

# The lawful approach?

Problems:

1. None is binding → GDPR?
2. Quantity over quality?
3. Too Western-centric?
4. Not democratic, neither participative

I wish to sign the Montreal Declaration
for a responsible development of artificial intelligence

I'm signing



## THE DECLARATION

# Agenda

1. **Intro**
    a. AI
    b. Ethics
2. **Ethical Implications**
    a. problems → values → actions
3. **Bias, A hands-on case study**
    a. The COMPASS algorithm

# Bias, a case study

1. **Bias, a definition**
    a. **Social Psychology, Law**
    b. **Statistics**
    c. **in algorithms?**
    d. **Case of racial bias**

# Bias, a case study

***Bias vs Discrimination vs Fairness***

# Bias, a case study

***Bias → Discrimination → Fairness?***

# Bias, a case study

***Bias → Discrimination → Fairness?***

# Bias, a case study

***Bias → Discrimination → Fairness?***

1. **Bias, a definition**
    a. **Social Psychology, Law**
    b. **Statistics**
    c. **in algorithms?**

# Bias, a case study

1. **Social Psychology, Law**
    a. **Negatives attitudes or prejudices towards a particular group**

Source:[1] Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), London. Nuffield Foundation

# Bias, a case study

1. **Social Psychology, Law**
    a. **Negatives attitudes or prejudices towards a particular group**
    b. **Implicit Bias**

Source:[1] Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), London. Nuffield Foundation

# Bias, a case study

*Implicit Association Test*

*Source: https://implicit.harvard.edu/implicit/takeatest.html*

Bi

| | |
|---|---|
| Asian IAT | ***Asian American*** **('Asian - European American' IAT).** This IAT requires the ability to recognize White and Asian-American faces, and images of places that are either American or Foreign in origin. |
| Weight IAT | ***Weight*** **('Fat - Thin' IAT).** This IAT requires the ability to distinguish faces of people who are obese and people who are thin. It often reveals an automatic preference for thin people relative to fat people. |
| Presidents IAT | ***Presidents*** **('Presidential Popularity' IAT).** This IAT requires the ability to recognize photos of Donald Trump and one or more previous presidents. |
| Skin-tone IAT | ***Skin-tone*** **('Light Skin - Dark Skin' IAT).** This IAT requires the ability to recognize light and dark-skinned faces. It often reveals an automatic preference for light-skin relative to dark-skin. |
| Race IAT | ***Race*** **('Black - White' IAT).** This IAT requires the ability to distinguish faces of European and African origin. It indicates that most Americans have an automatic preference for white over black. |
| Disability IAT | ***Disability*** **('Disabled - Abled' IAT).** This IAT requires the ability to recognize symbols representing abled and disabled individuals. |
| Gender-Science IAT | ***Gender - Science.*** This IAT often reveals a relative link between liberal arts and females and between science and males. |
| Weapons IAT | ***Weapons*** **('Weapons - Harmless Objects' IAT).** This IAT requires the ability to recognize White and Black faces, and images of weapons or harmless objects. |
| Arab-Muslim IAT | ***Arab-Muslim*** **('Arab Muslim - Other People' IAT).** This IAT requires the ability to distinguish names that are likely to belong to Arab-Muslims versus people of other nationalities or religions. |
| Transgender IAT | ***Transgender*** **('Transgender People – Cisgender People' IAT).** This IAT requires the ability to distinguish photos of transgender celebrity faces from photos of cisgender celebrity faces. |
| Sexuality IAT | ***Sexuality*** **('Gay - Straight' IAT).** This IAT requires the ability to distinguish words and symbols representing gay and straight people. It often reveals an automatic preference for straight relative to |

# Bias, a case study

1. **Statistics**
    a. **a ‘biased sample’ means a sample that does not adequately represent the distribution of features in the reference population (e.g. it contains a higher proportion of young men than in the overall population).**

Source:[1] Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Whittlestone, J. Nyrup, R. Alexandrova, A. Dihal, K. Cave, S. (2019), London. Nuffield Foundation

# Bias, a case study

1. **Bias in Machine Learning**
    a. **errors in estimation or over/under representing populations when sampling.**
    b. **i.e. strawberries & melons (by Prof. Zicari)**

Source: CS 294: Fairness in Machine Learning UC Berkeley, Fall 2017 https://mrtz.org/nips17/#/6

# Bias, a case study

**1. Strawberries & Melons**

***Two people are tasked with developing a system to sort a basket of fruit.***

***They have to determine which pieces are “high quality” and will be sold at the market, and which will instead be used for making jam. “***

Source: Understanding Bias in Algorithmic Design https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

# Bias, a case study

## 1. Strawberries & Melons

***Two people are tasked with developing a system to sort a basket of fruit.***

***They have to determine which pieces are "high quality" and will be sold at the market, and which will instead be used for making jam. "***

***Your have strawberries, melon and bananas. What would be your criteria for "quality"? Which ones will you send to the market?***

Source: Understanding Bias in Algorithmic Design https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

# Bias, a case study

1. **Strawberries & Melons**
   a. ***Person 1: Quality = Brightness of colour→ Strawberries to the market, melons to the jam factory***

   b. ***Person 2: Quality = When Unblemished → Ripe melons & bananas to the market, unripe strawberries to the jam factory***

Source: Understanding Bias in Algorithmic Design https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

# Bias, a case study

1. **Lesson:**
   a. ***'Similarly logical & evenly applied criteria will result in 2 different outcomes for the same basket of fruit'***
   b. **Ok for fruit…**
   c. **… But for humans?**

Source: Understanding Bias in Algorithmic Design https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

# Bias, a case study

1. **For humans?**
    a. **ML algorithms are used to automatize or assist decisions**
        1. **Healthcare: diagnosis, treatment**
        2. **Judiciary: Recidivism risk assessments**
        3. **Recruiting: Screening job applicants**
        4. **Journalism: News Recommender Systems**
        5. **Banking: Credit ratings/Loan approvals**
        6. **Welfare: Welfare Benefit Eligibility**

# Bias, a case study

1. **For humans?**
    a. **In which cases of Ai application can bias arise**
        1. **Healthcare: diagnosis, treatment**
        2. **Judiciary: Recidivism risk assessments**
        3. **Recruiting: Screening job applicants**
        4. **Journalism: News Recommender Systems**
        5. **Banking: Credit ratings/Loan approvals**
        6. **Welfare: Welfare Benefit Eligibility**

# Bias, a case study

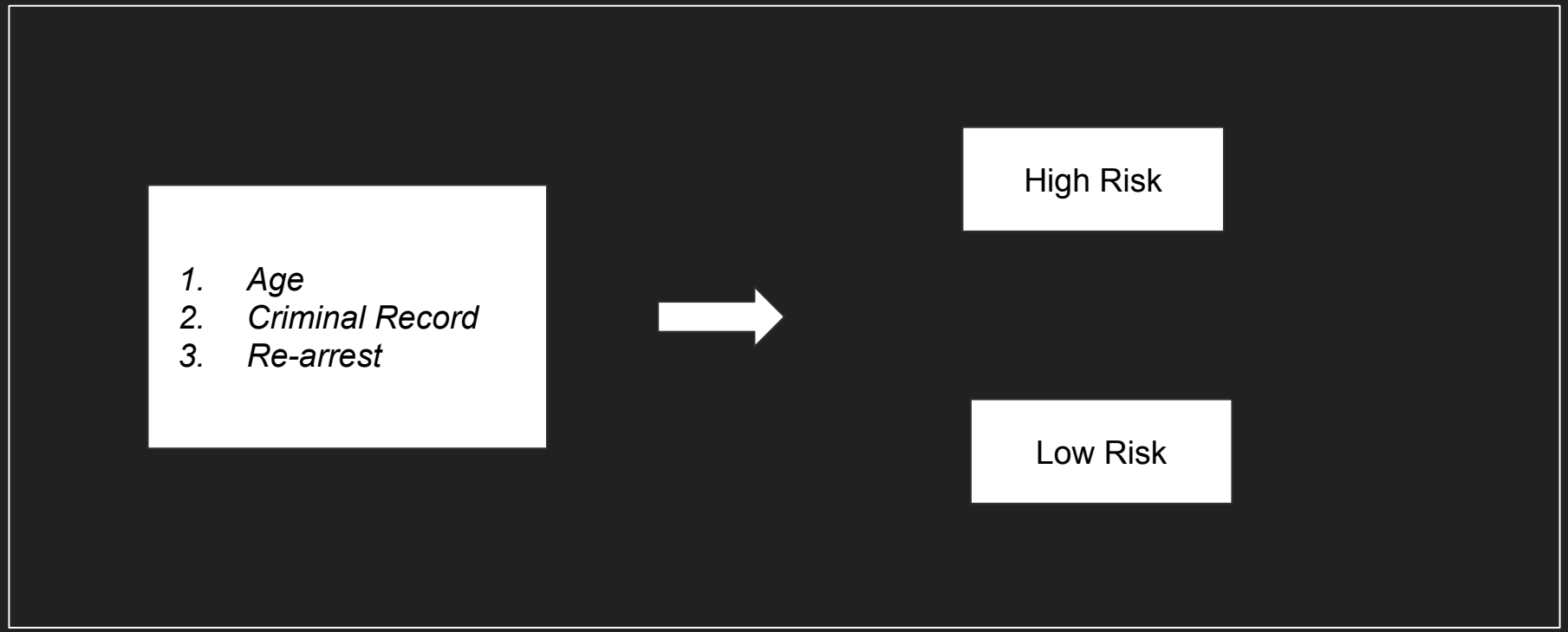
1. **For humans?**
    a. **In which cases of Ai application can bias arise**
        1. **Healthcare: diagnosis, treatment**
        2. **Judiciary: Recidivism risk assessments**
        3. **Recruiting: Screening job applicants**
        4. **Journalism: News Recommender Systems**
        5. **Banking: Credit ratings/Loan approvals**
        6. **Welfare: Welfare Benefit Eligibility**

# Bias, a case study

1. **The case of racial bias**
    a. **COMPASS, software used in the US to predict the likelihood of a person committing a future crime**
    b. **Racial bias**
        i. **A study by Angwin and colleagues in 2016 found that it is biased towards people of colour**

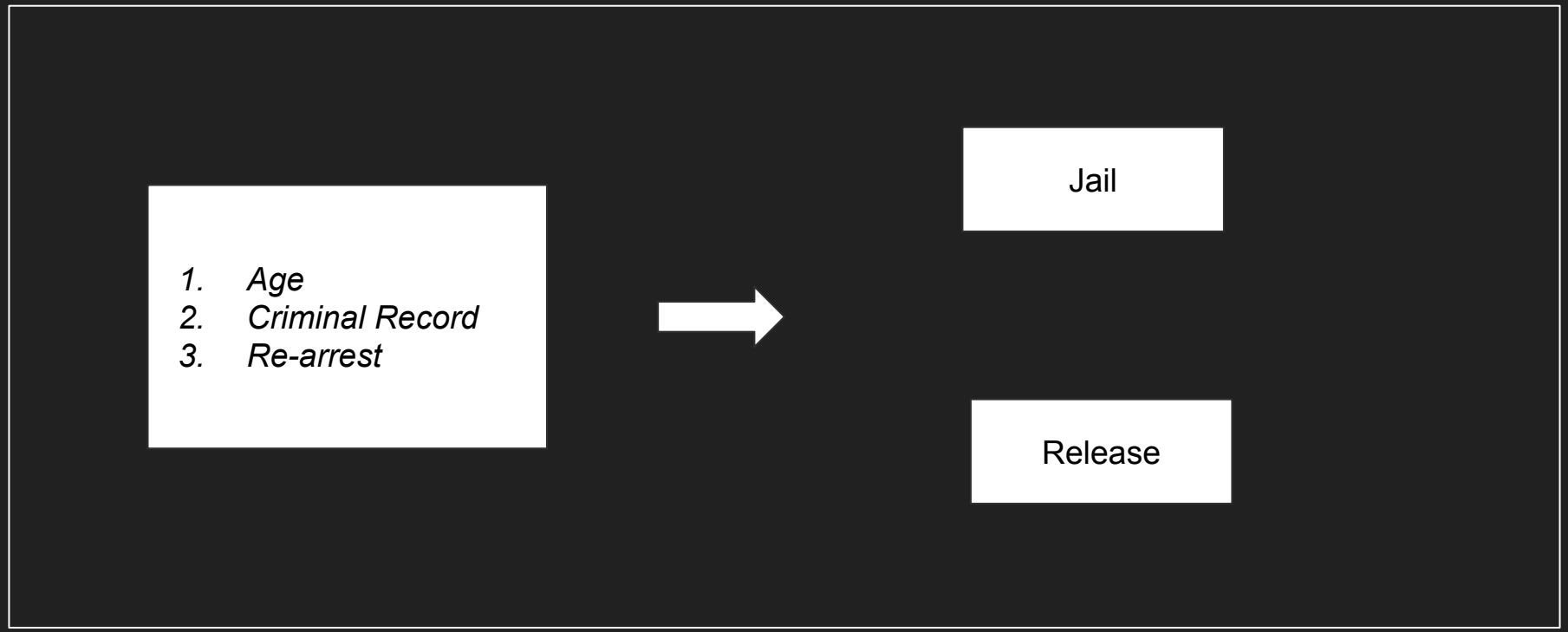
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study



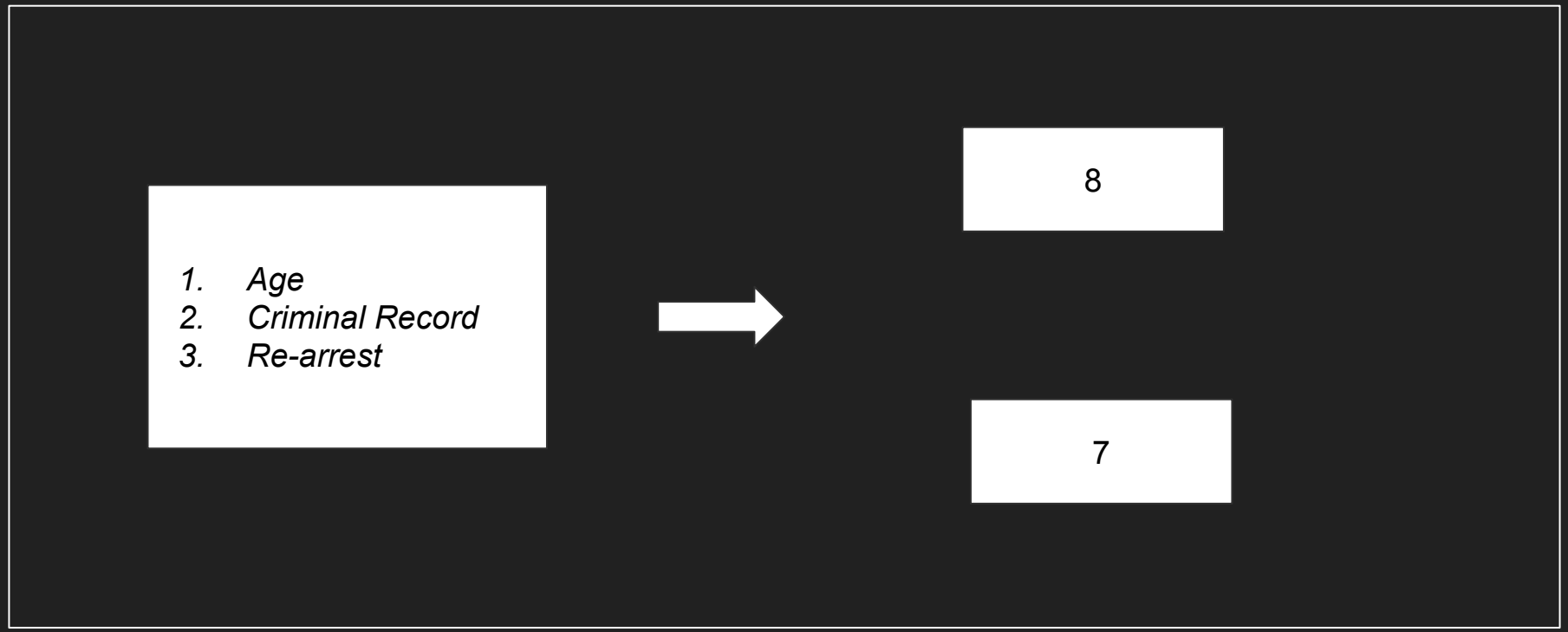

Source: http://htmlpreview.github.io/?https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

# Bias, a case study




Source: http://htmlpreview.github.io/?https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

# Bias, a case study



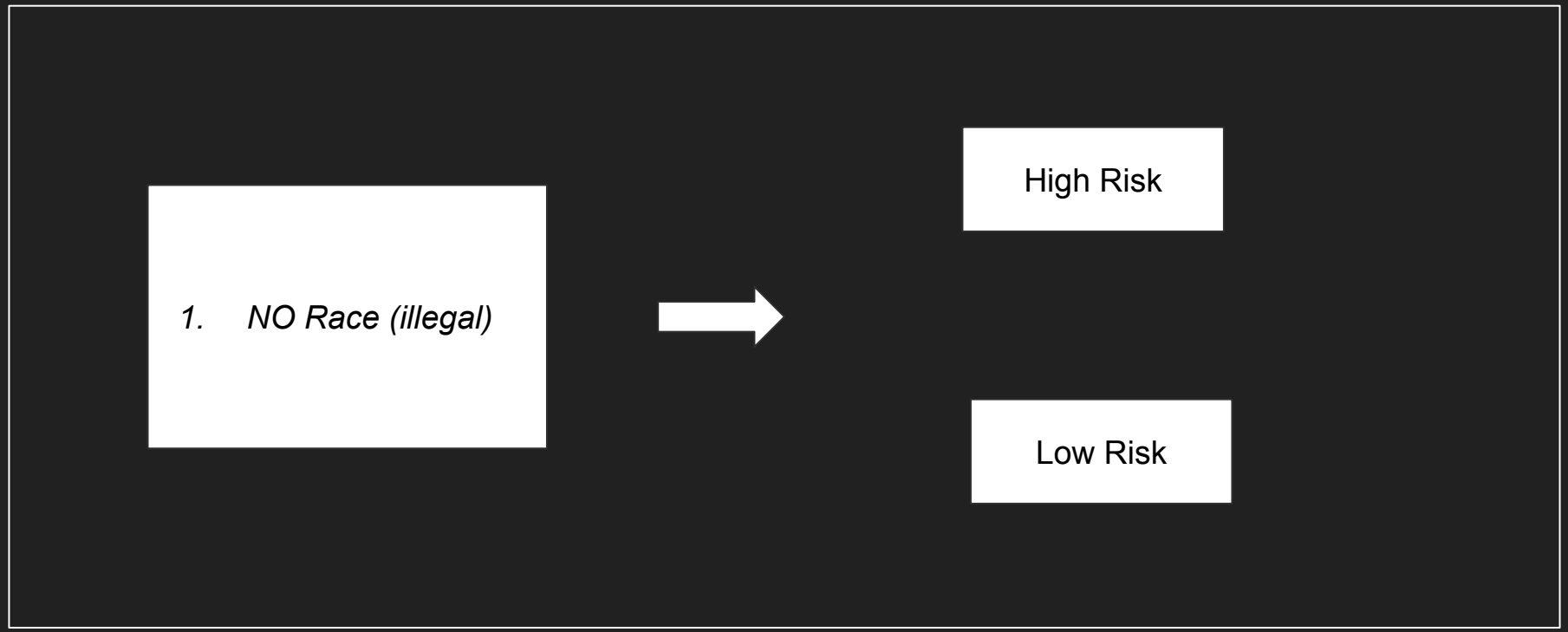

Source: http://htmlpreview.github.io/?https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

# Bias, a case study

1. **The case of racial bias**
    a. **Arguments for pretrial risk-assessment tools**
        i. **Eliminates Judges' Bias**
        ii. **No more posting bail (money)**

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study




Source: http://htmlpreview.github.io/?https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

# Bias, a case study

1. **The case of racial bias**
    a. **Arguments for pretrial risk-assessment tools**
        i. **Eliminates Judges' Bias**
        ii. **No more posting bail (money)**
    b. **Arguments against**
        i. **Even though race cannot be used as a variable…**
            1. **It still discriminates**

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study

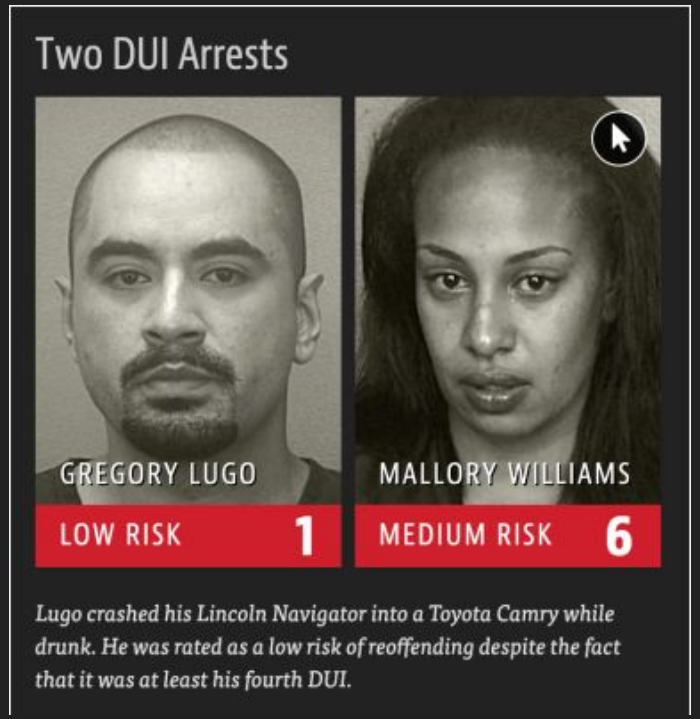
1. **The case of racial bias**
    a. **Arguments for pretrial risk-assessment tools**
        i. **Eliminates Judges' Bias**
        ii. **No more posting bail (money)**
    b. **Arguments against**
        i. **Even though race cannot be used as a variable…**
            1. **It still discriminates**

**WHY?**

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing


# Bias, a case study

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing




# Bias, a case study

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing


# Bias, a case study

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

## Two Drug Possession Arrests


*Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.*

# Bias, a case study

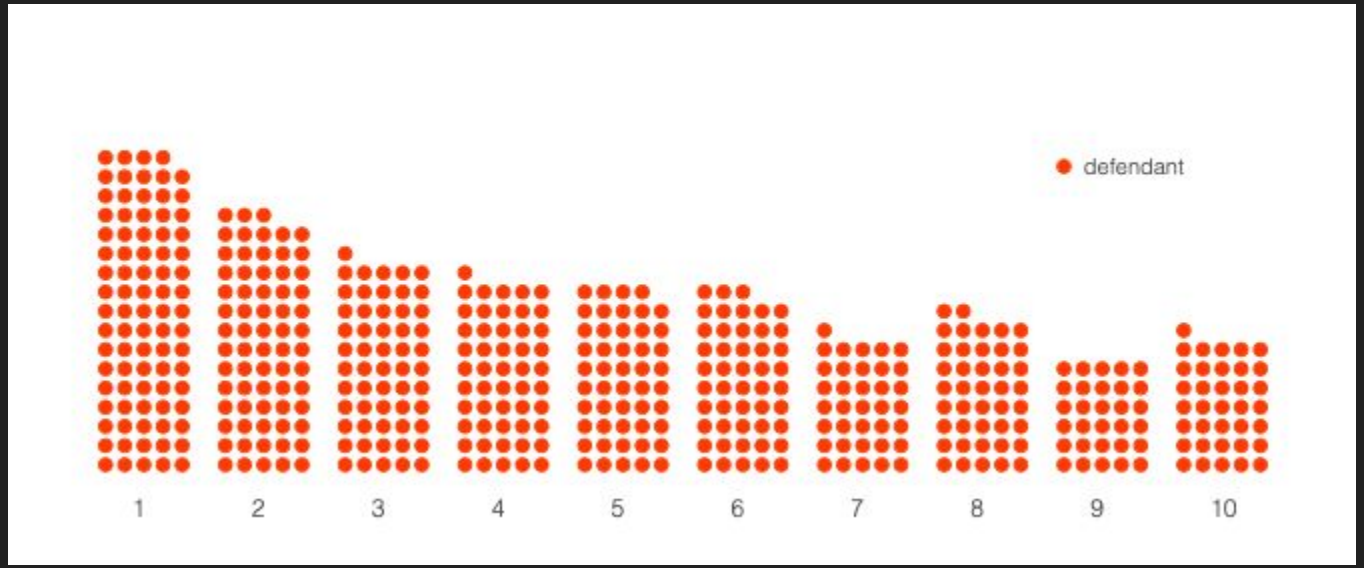
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

1. **COMPASS**
   a. **Try it out yourself: ‘Can you make AI fairer than a judge?’ article**
   b. **Source:** https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

# Let’s try to make it fair!

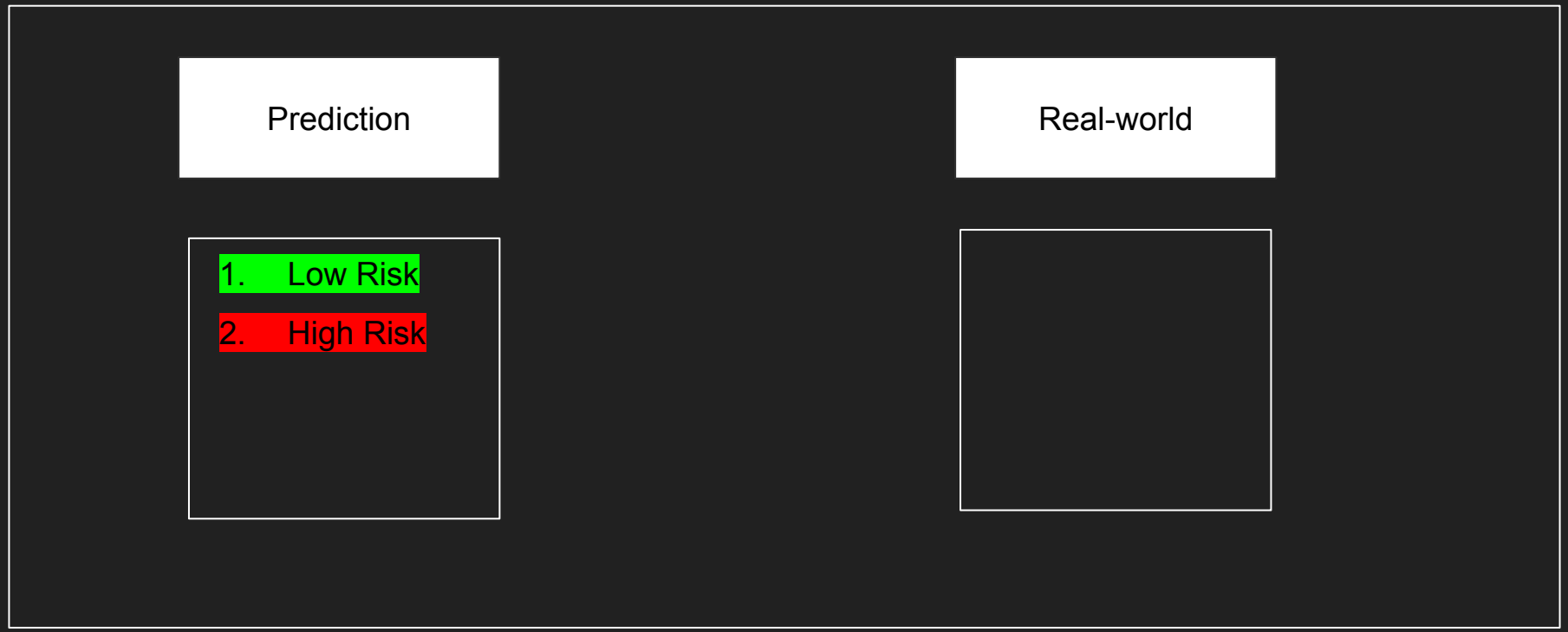
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study



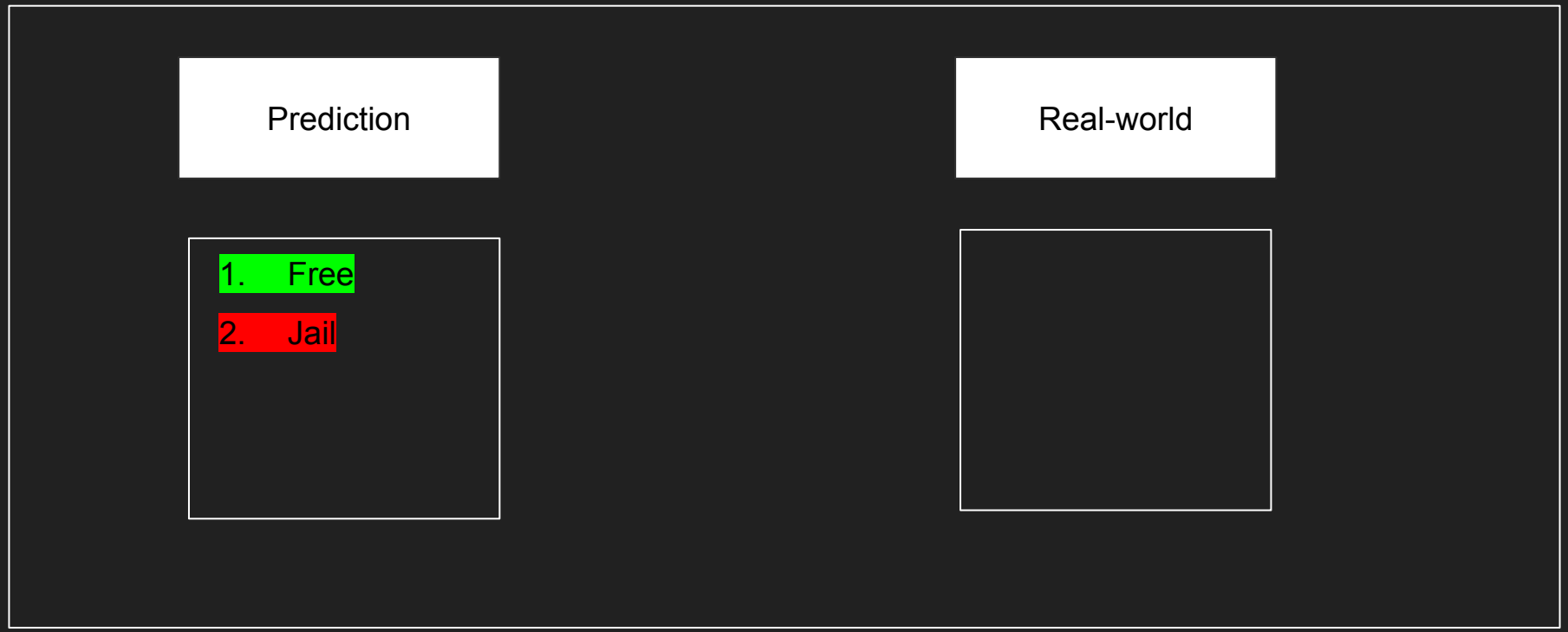

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study




Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study



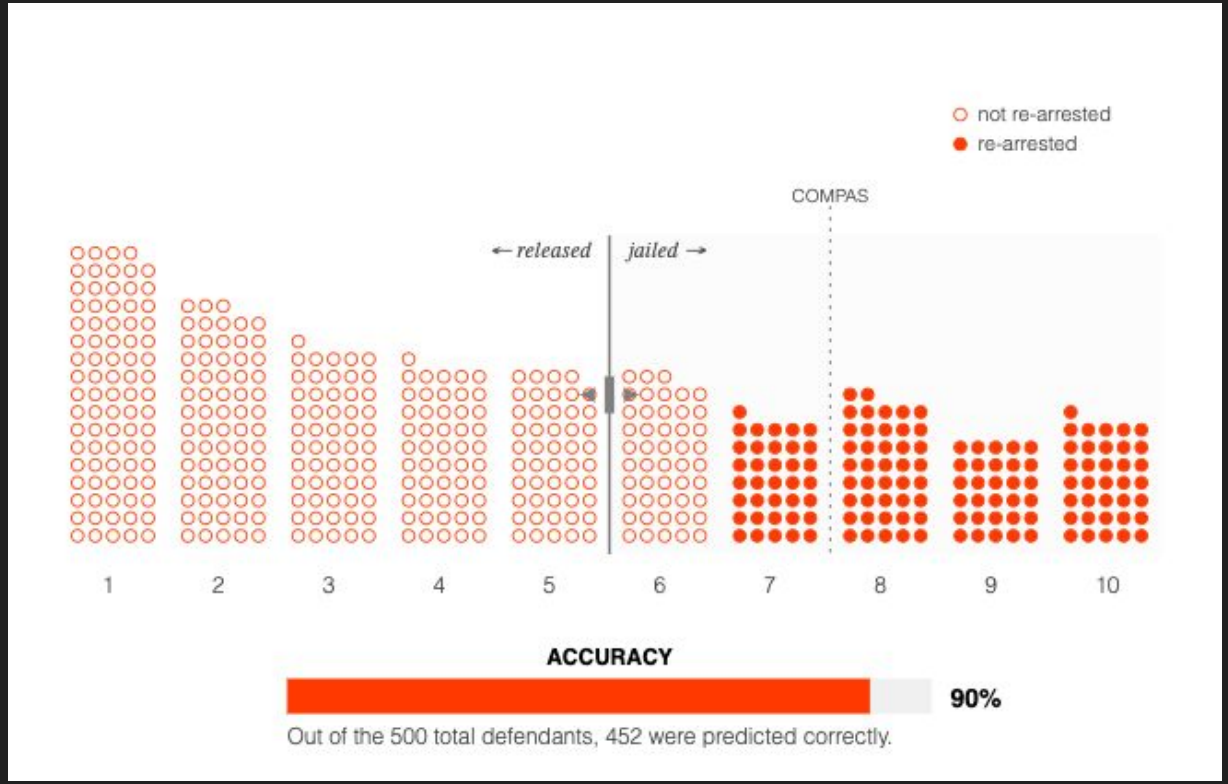

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study

| Prediction | Real-world |
| --- | --- |
| 1. Low Risk | 1. No re-arrest |
| 2. High Risk | 2. Re-arrest |

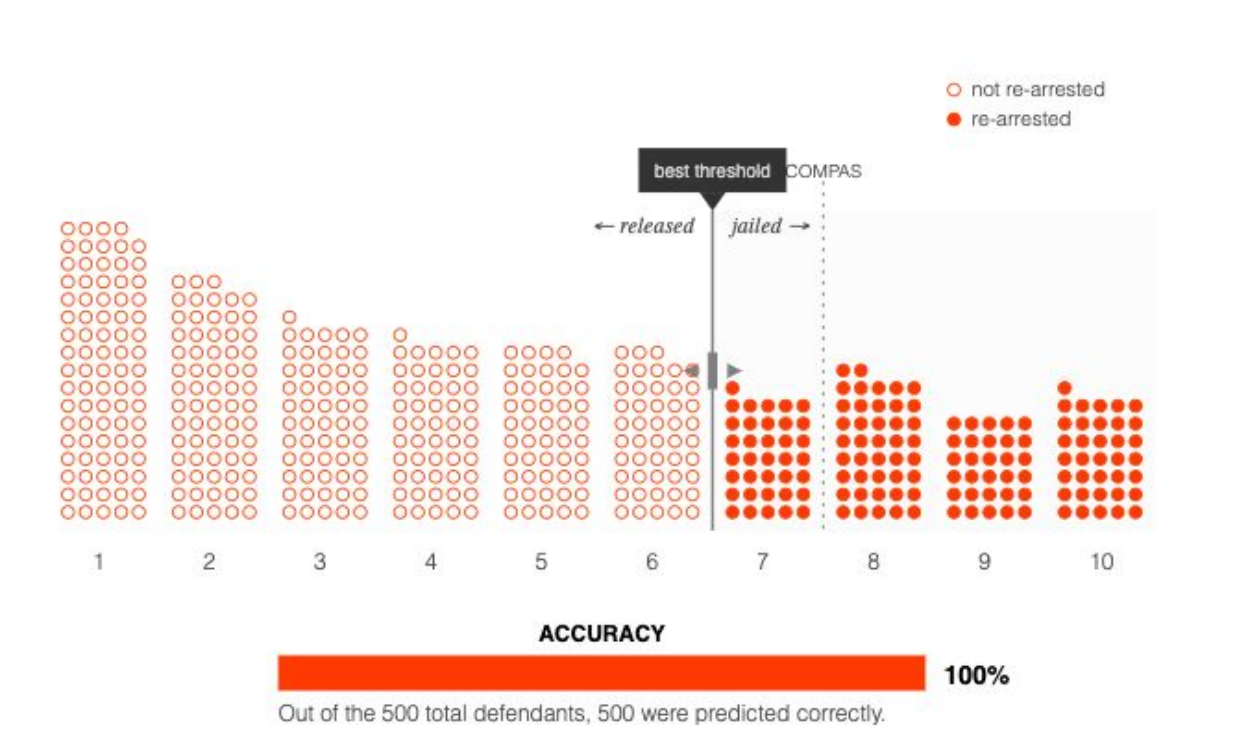
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study



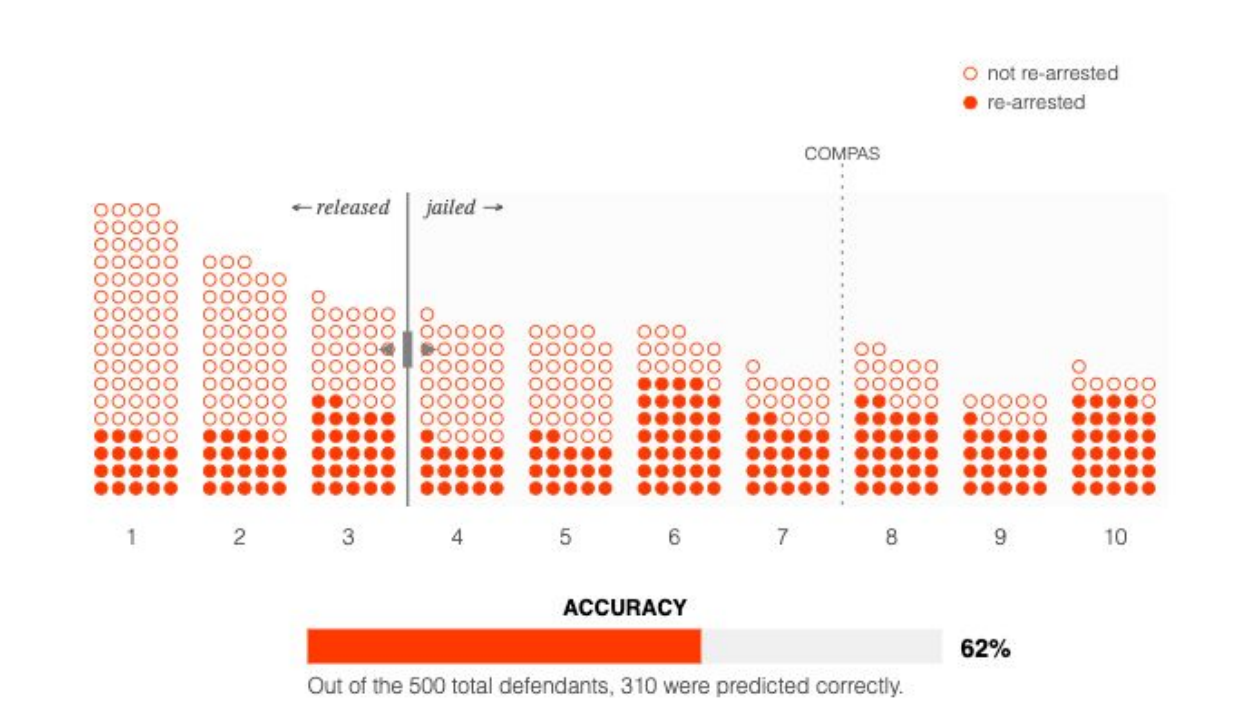

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study




Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study




Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

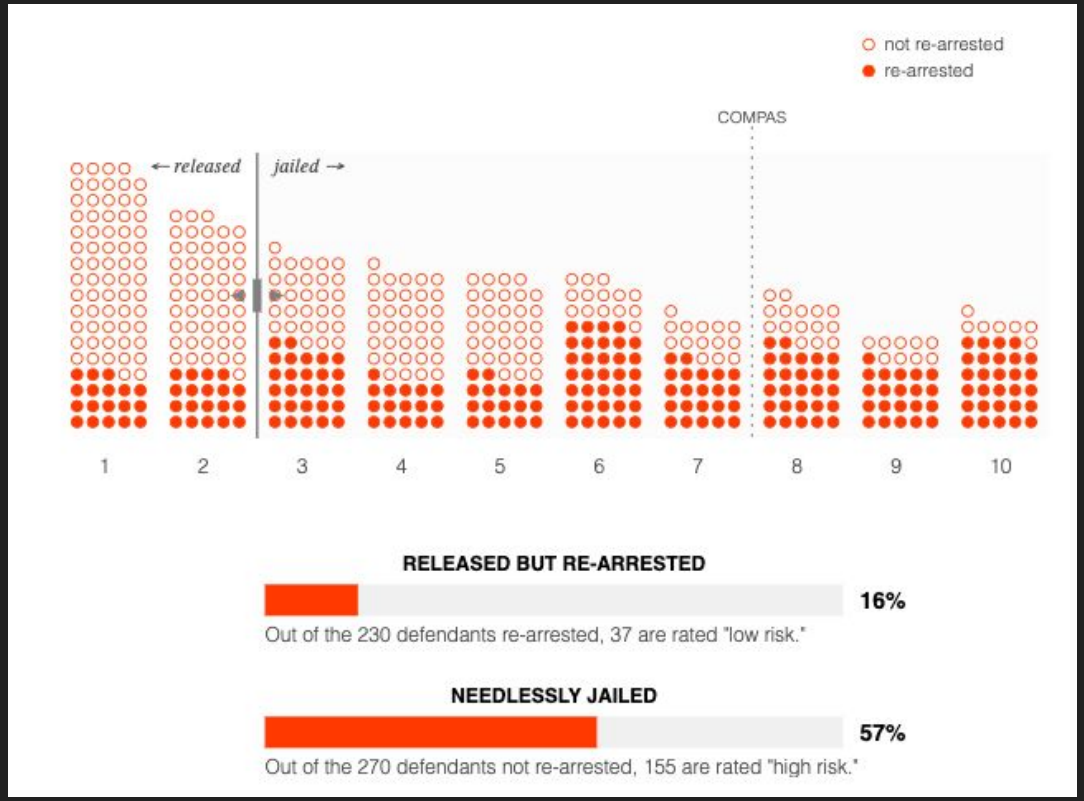
# Bias, a case study

*“It is better that 10 guilty persons escape than that one innocent suffer”*,

William Blackstone

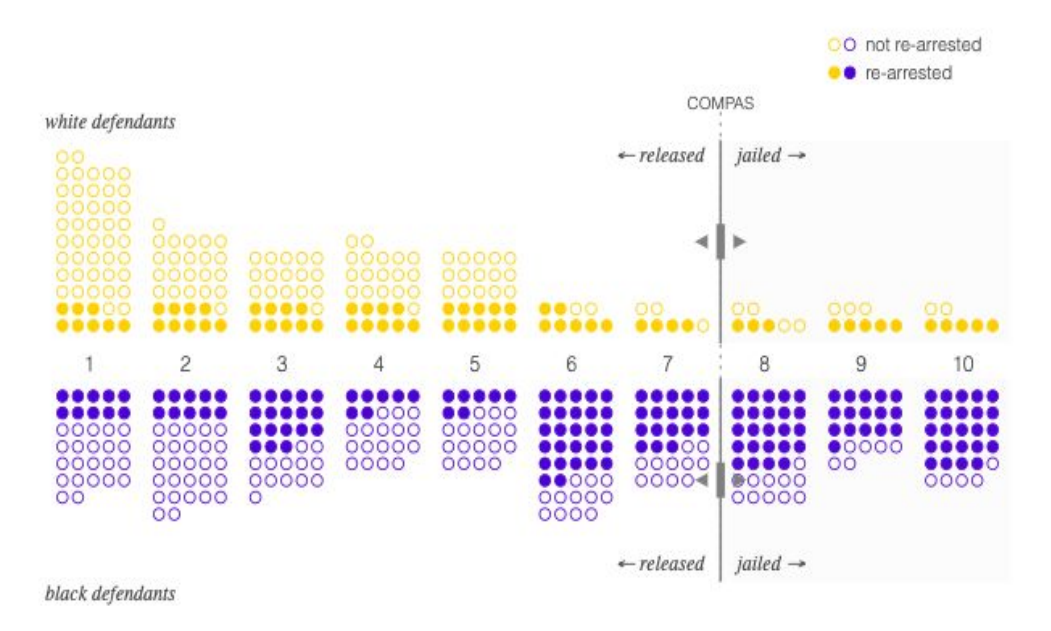
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study




Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

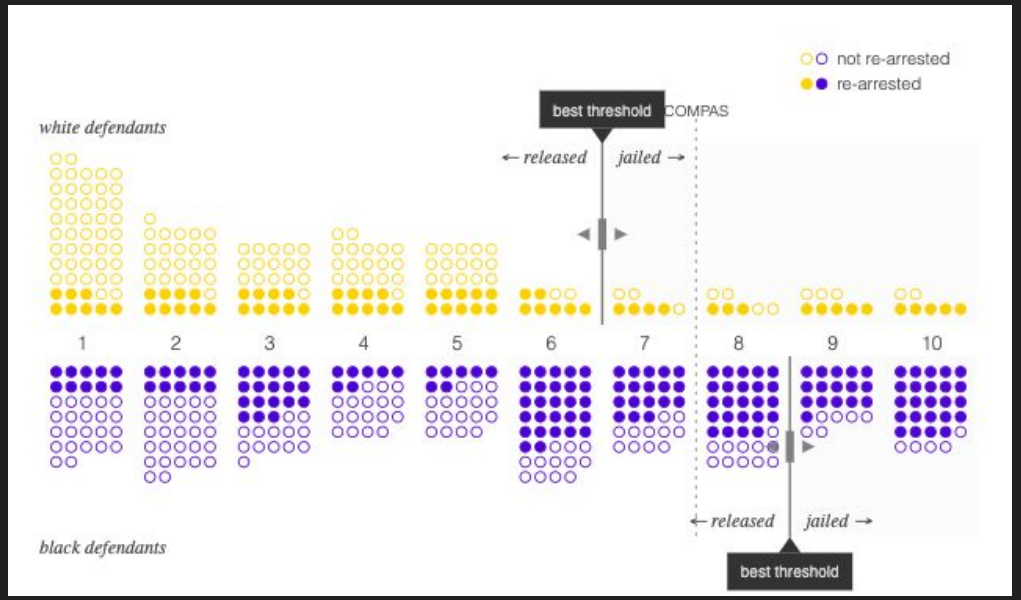
# Bias, a case study





Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study





Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

1. Mmh.. no?
    a. **Fairness 1:** treat people with the same risk scores in the same way
    b. **Fairness 2:** keep error rates comparable between groups

# Is there a solution?

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Bias, a case study

1. **For humans?**
    a. **In which cases of Ai application can bias arise**
        1. **Healthcare: diagnosis, treatment**
        2. **Judiciary: Recidivism risk assessments**
        3. **Recruiting: Screening job applicants**
        4. **Journalism: News Recommender Systems**
        5. **Banking: Credit ratings/Loan approvals**
        6. **Welfare: Welfare Benefit Eligibility**

1. **COMPASS**

a. “Though judges may not always be transparent about how they choose between different notions of fairness, people can contest their decisions”

# What has changed?

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

1. **COMPASS**

   a. In contrast, COMPAS, which is made by the private company Northpointe, is a trade secret that cannot be publicly reviewed or interrogated.

**What has changed?**

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Thank you!


Marta:

*Email:* marta.ziosi@aiforpeople.org

AI for People:

*Website*: www.aiforpeople.org

LinkedIn:linkedin.com/company/ai-for-people/

*Twitter:* https://twitter.com/AIforPeople

*Facebook*: www.facebook.com/aiforpeople

*Medium*: www.medium.com/ai-for-people