Ethical AI

HENNELLY & GROSSFELD LLP

MIKE KING

4640 ADMIRALTY WAY
SUITE 850
MARINA DEL REY, CA 90292
(310) 305-2100
www.hgla.com

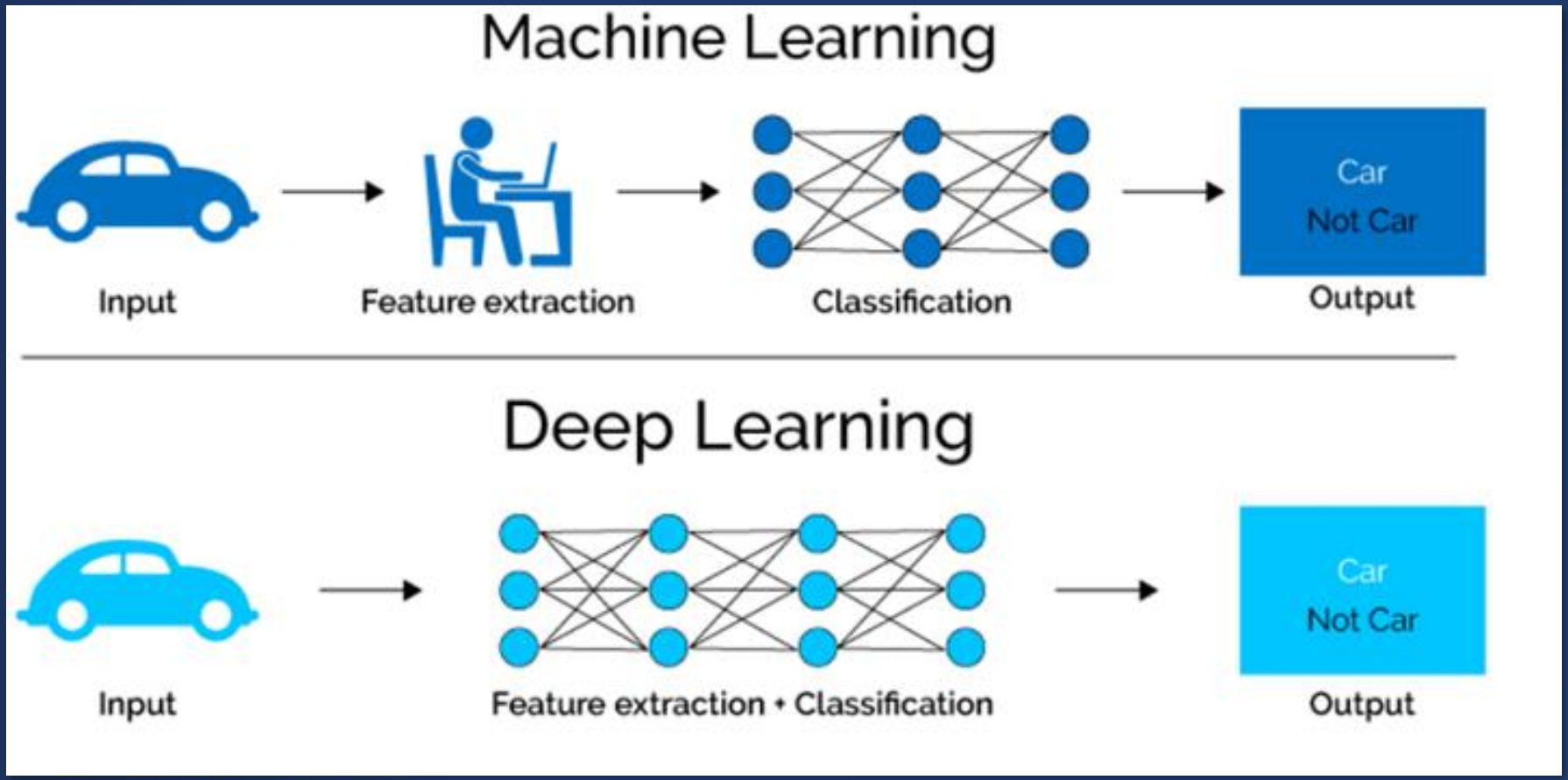
Isaac Asimov's Three Laws of Robotics (1940)

First Law: **A robot may not injure a human or through inaction, allow a human to come to harm.**

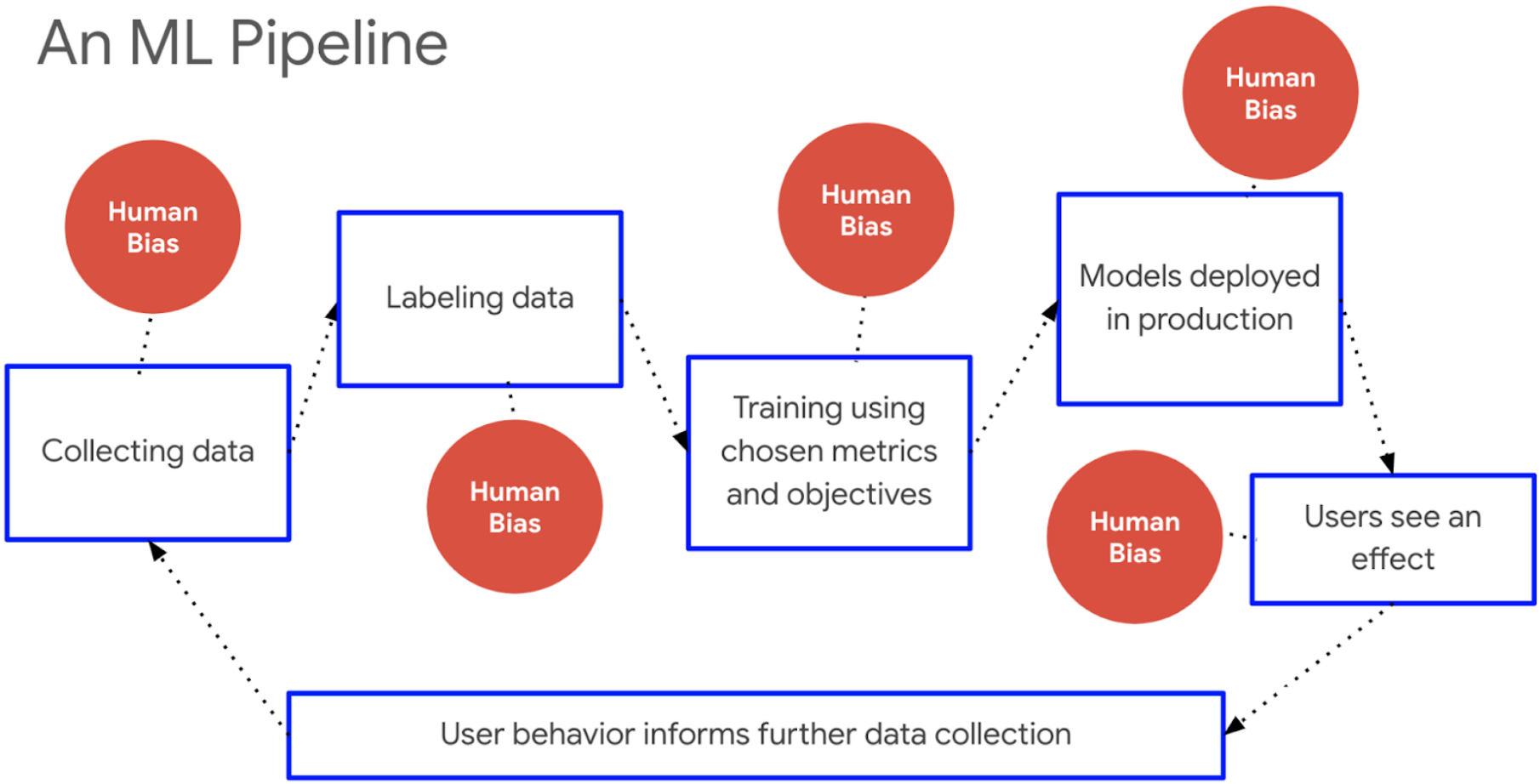
Second Law: **A robot must obey the orders given it by human beings, unless such orders would conflict with the first law.**

Third Law: **A robot must protect its own existence, as long as such protection does not conflict with the first or second law.**

AI can be described as a set of math functions (**model**) that given some inputs (**data**), learn *something* and use that to *infer* something else (make **predictions**). In other words AI is data, model and predictions. Ethical exploration of this realm covers issues like **bias** in a models' predictions and **fairness** (or lack thereof) of the outcomes; as well as approaches to address them via **accountability** and **transparency**.


Machine Learning
Input
Feature extraction
Classification
Car
Not Car
Output
Deep Learning
Input
Feature extraction + Classification
Car
Not Car
Output

An ML Pipeline




AI Fairness 360 Toolkit (AIF360)

FAST COMPANY

10.28.19 | 8:00 AM

Technology biased against black patients runs rampant in hospitals

A new study shows that a widely used algorithm for predicting which patients get additional care is disproportionately counting out black patients—and could have left tens of thousands without adequate medical care.

[Photo: Image Source/Getty Images]

5 sources of bias: Historical bias *already* exists in the data while representation bias and measurement bias are a result of how the dataset is created. Evaluation and aggregation biases are a result of the choices made while building the model.

"Unfortunately, we have biases that live in our data, and if we don't acknowledge that and if we don't take specific actions to address it then we're just going to continue to perpetuate them or even make them worse."

— Kathy Baxter, Ethical AI Practice Architect, Salesforce

"There's a real *danger of systematizing the discrimination* we have in society [through AI technologies]. What I think we need to do — as we're moving into this world full of invisible algorithms everywhere — is that we have to be very explicit, or have a disclaimer, about what our ***error rates*** are like."

— Timnit Gebru, Research Scientist, Google AI

“There is a *silver lining* on the bias issue. For example, say you have an algorithm trying to predict who should get a promotion. And say there was a supermarket chain that, statistically speaking, didn’t promote women as often as men. It *might be easier to fix an algorithm than fix the minds of 10,000 store managers*.”

— Richard Socher, Chief Scientist, Salesforce

"We're seeing a kind of a *Wild West* situation with *AI and regulation* right now. The scale at which businesses are adopting AI technologies isn't matched by clear guidelines to regulate algorithms and help researchers avoid the pitfalls of bias in datasets. We need to advocate for a *better system of checks and balances to test AI for bias and fairness*, and to help businesses determine whether certain use cases are even appropriate for this technology at the moment."

— Timnit Gebru, Research Scientist, Google AI

AI & Law
Bail, Sentencing & Parole

AI & Law

Q: COMPAS violate due process b/c proprietary nature hinders challenging scientific validity or because it takes gender into account?

A: No.

Warning required

Judge must independently determine

State of Wisconsin v. Loomis, 881 N.W.2d 749 (2016)

COMPAS & Bias

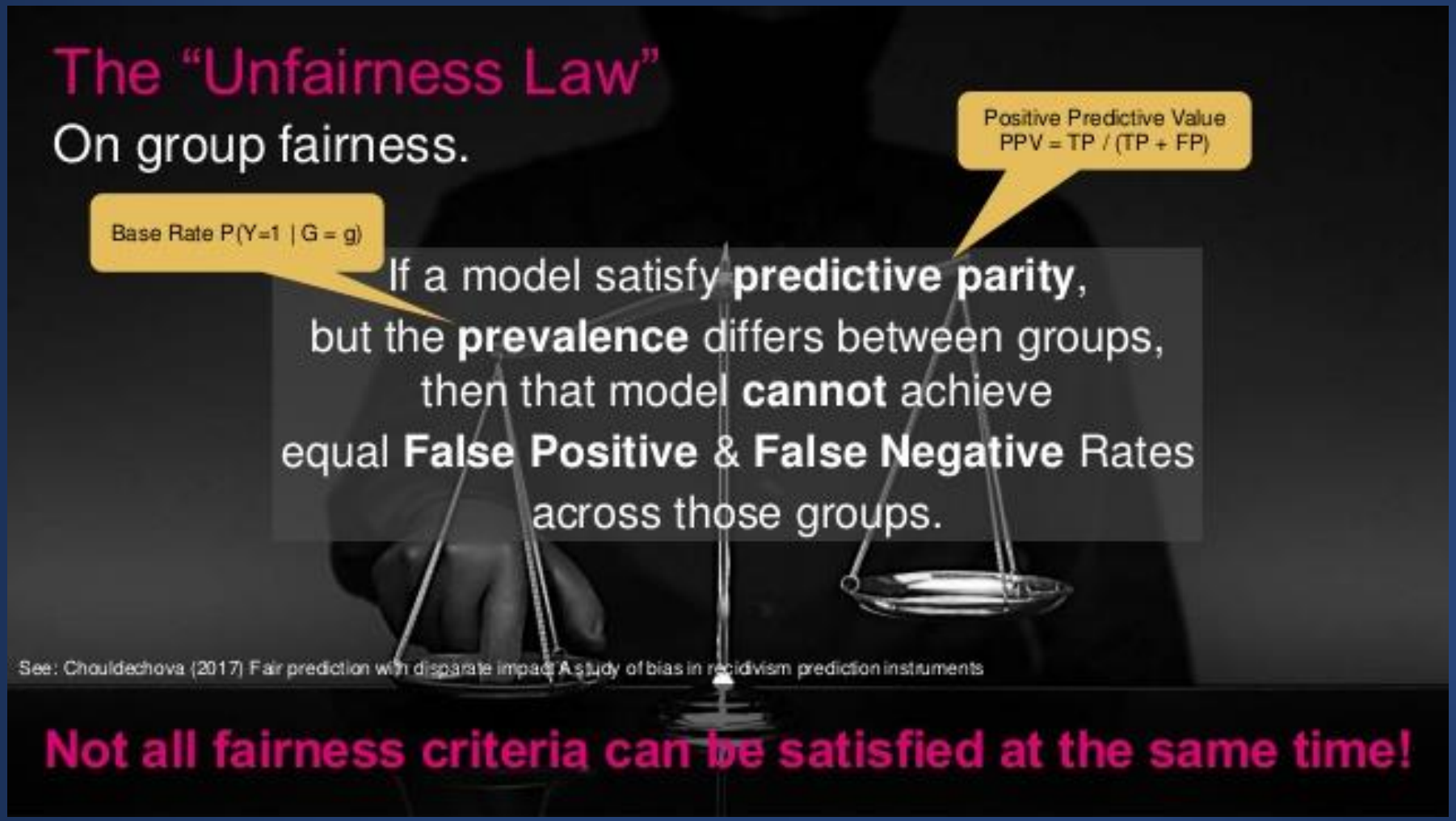
ProPublica claimed COMPAS program was biased using “false positives” analysis

Black defendants who did not recidivate within 2 years were nearly twice as likely to be misclassified as higher risk compared to white defendants (45% v. 23%)

COMPAS & Bias

COMPAS claimed program was not biased because it satisfied “predictive parity”

COMPAS scores accuracy rate was the same for black and white defendants (about 60%)


The "Unfairness Law"
On group fairness.
Positive Predictive Value
PPV = TP / (TP + FP)
Base Rate P(Y=1 | G = g)
If a model satisfy **predictive parity**,
but the **prevalence** differs between groups,
then that model **cannot** achieve
equal **False Positive** & **False Negative** Rates
across those groups.
See: Chouldechova (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments
Not all fairness criteria can be satisfied at the same time!

COMPAS & Bias

Existence of bias depended upon statistical measuring stick used:

false positives vs. predictive parity

EU's Ethical AI Guidelines

EU's Ethical AI Guidelines

EU's Ethical AI Guidelines

"Trustworthy AI has three components:

(1) it should be lawful, ensuring compliance with all applicable laws and regulations,

(2) it should be *ethical, ensuring adherence to ethical principles and values* and

(3) it should be *robust*, both from a technical and *social perspective* since to ensure that, even with good intentions, AI systems do not cause any unintentional harm."

EU's Ethical AI Guidelines

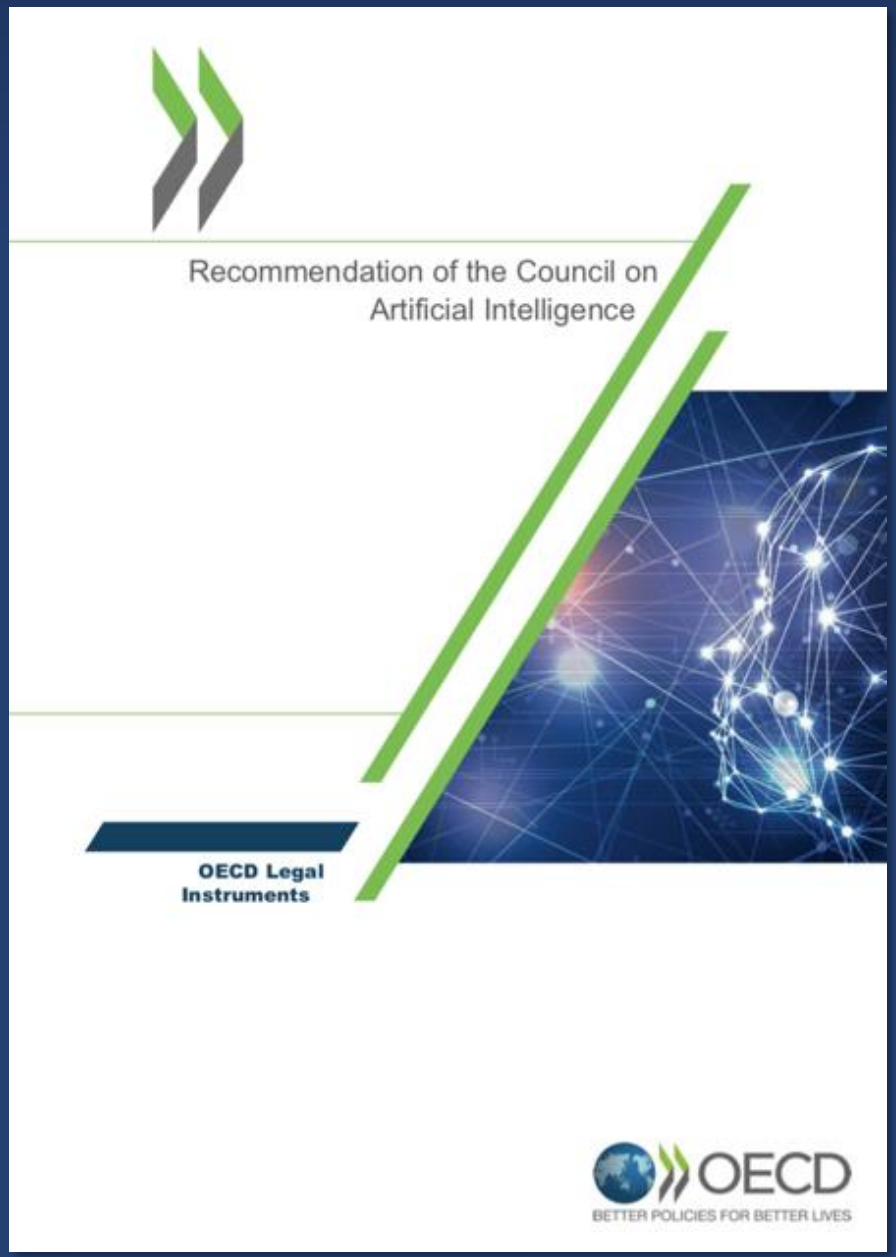
- Unfair bias avoidance: Did you establish a strategy or a set of procedures to *avoid creating or reinforcing unfair bias* in the AI system, both *input data and algorithm design*?

- Depending on the use case, did you ensure a mechanism that *allows others to flag issues related to bias, discrimination or poor performance* of the AI system?

EU's Ethical AI Guidelines

- Did you ensure an adequate working *definition of "fairness"* that you apply in designing AI systems?

OECD's AI Principles



OECD's Principles on AI

The OECD Artificial Intelligence (AI) Principles in short

- 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 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 when they are engaging with them and can challenge outcomes.
- AI systems must function in a robust, secure and safe way throughout their life cycle 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.

OECD

www.oecd.org/going-digital/ai

going digital

AI Organizations

MIT AI POLICY CONGRESS

AI NOW INSTITUTE

Stanford
Human-Centered
Artificial Intelligence

AI Organizations

Asilomar AI principles

RESEARCH

1. Research goal
2. Research funding
3. Science-policy link
4. Research culture
5. Race avoidance

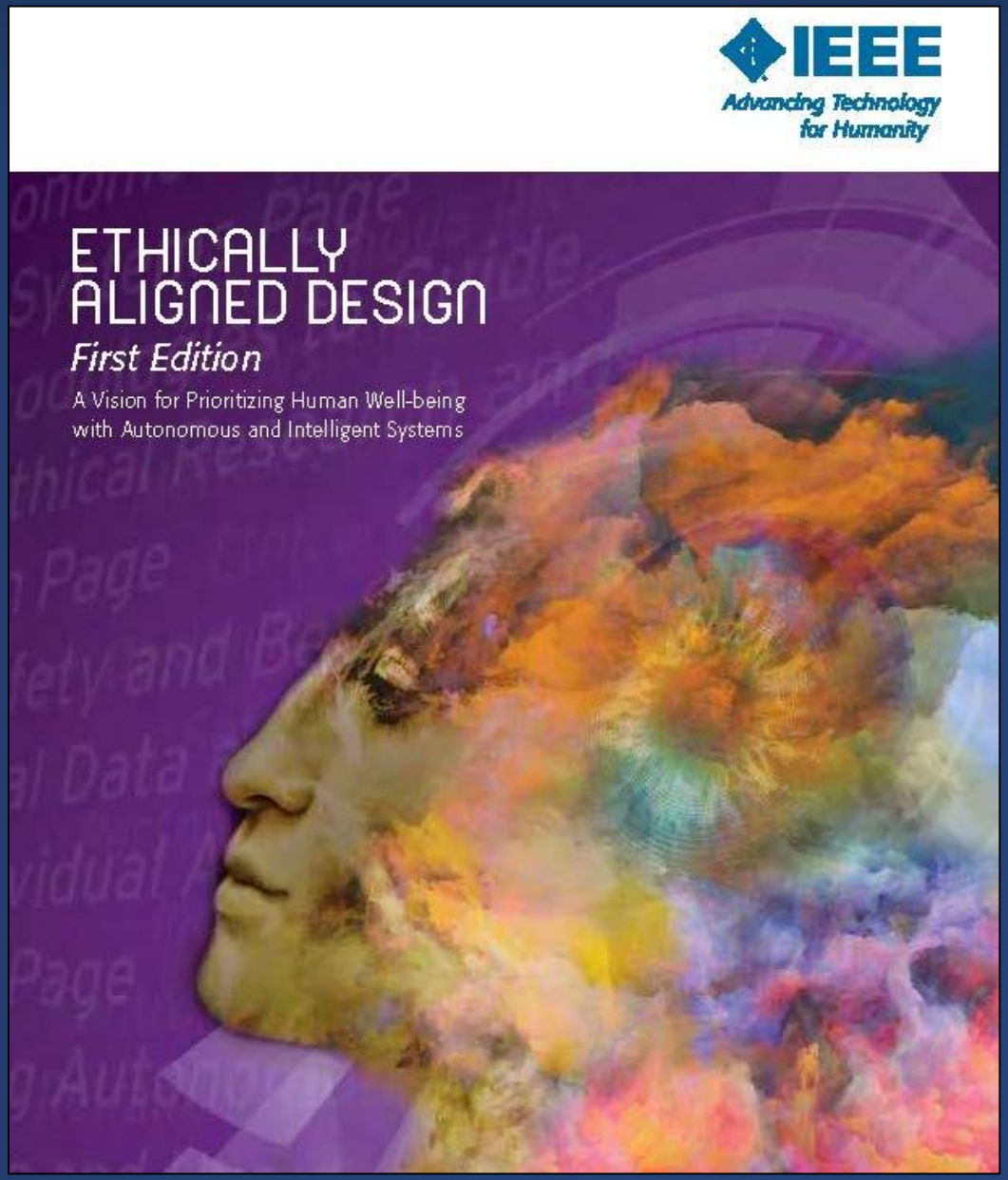
ETHICS AND VALUES

6. Safety
7. Failure transparency
8. Judicial transparency
9. Responsibility
10. Value alignmnet
11. Human values
12. Personal privacy
13. Liberty and privacy
14. Shared benefit
15. Shared prosperity
16. Human control
17. Non-subversion
18. AI arms race

LONGER-TERM ISSUES

19. Capability caution
20. Importance
21. Risks
22. Recursive self-improvement
23. Common good

IEEE AI Ethics



Microsoft's AI Principles

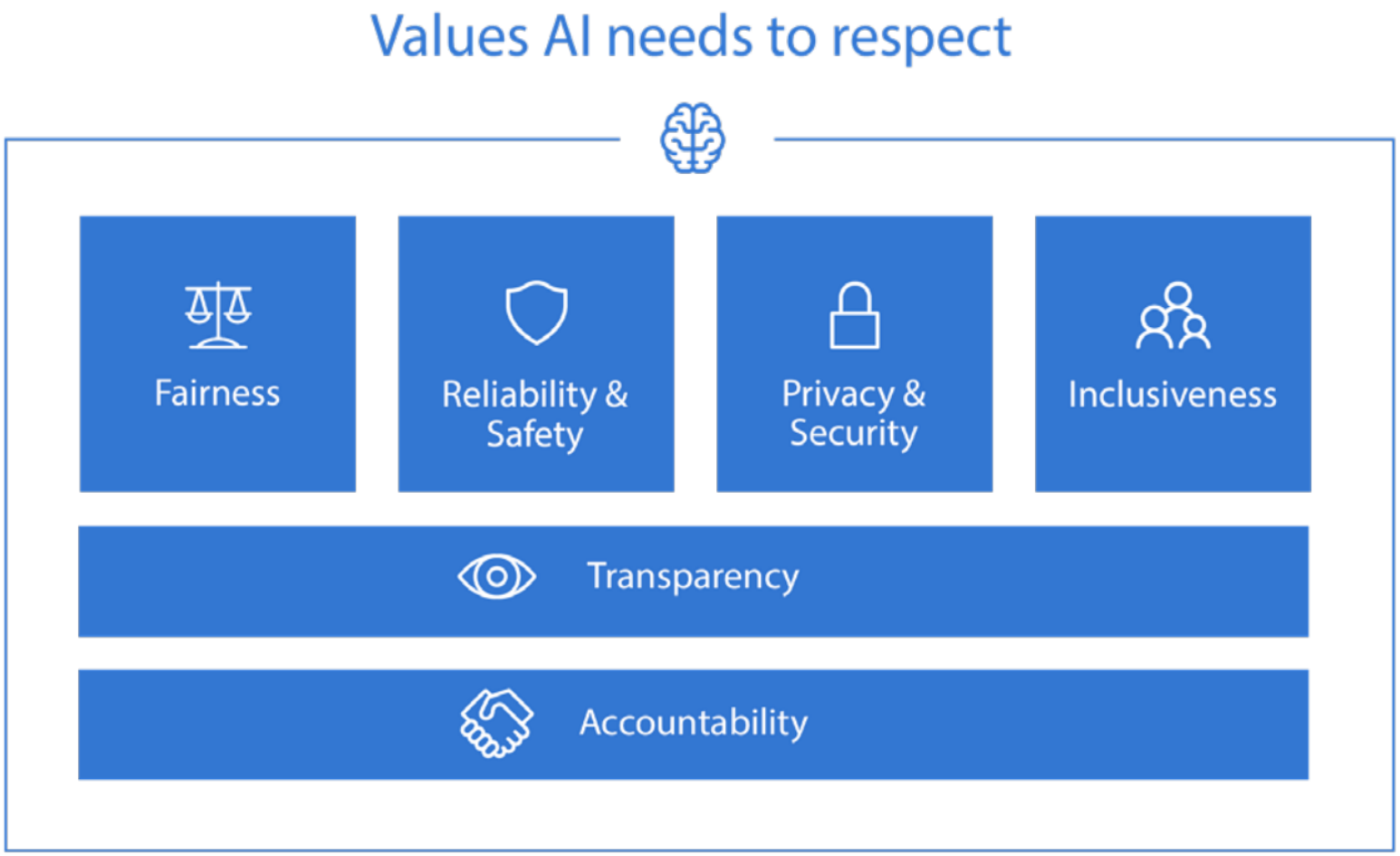


Chart 5.
Source: Microsoft Corporation

Microsoft's AI Principles

Microsoft will be adding AI ethics to its standard checklist for product release

BY ALAN BOYLE on March 25, 2019 at 2:09 pm

BOT or NOT? This special series explores the evolving relationship between humans and machines, examining the ways that robots, artificial intelligence and automation are impacting our work and lives.

WSJ PRO
ARTIFICIAL INTELLIGENCE

Vatican Advisory Group Issues Call for AI Ethics

IBM and Microsoft have signed on to the Pontifical Academy for Life's charter on artificial intelligence

By *John McCormick*

Feb. 28, 2020 7:30 am ET | WSJ PRO

Recent Developments

Perspectives on Issues in AI Governance

AI & Law

"…this paper is a call for governments and civil society groups worldwide to make a substantive contribution to the AI governance discussion. Specifically… explainability standards, ***approaches to appraising fairness***, safety considerations, requirements for human-AI collaboration, and general liability frameworks."

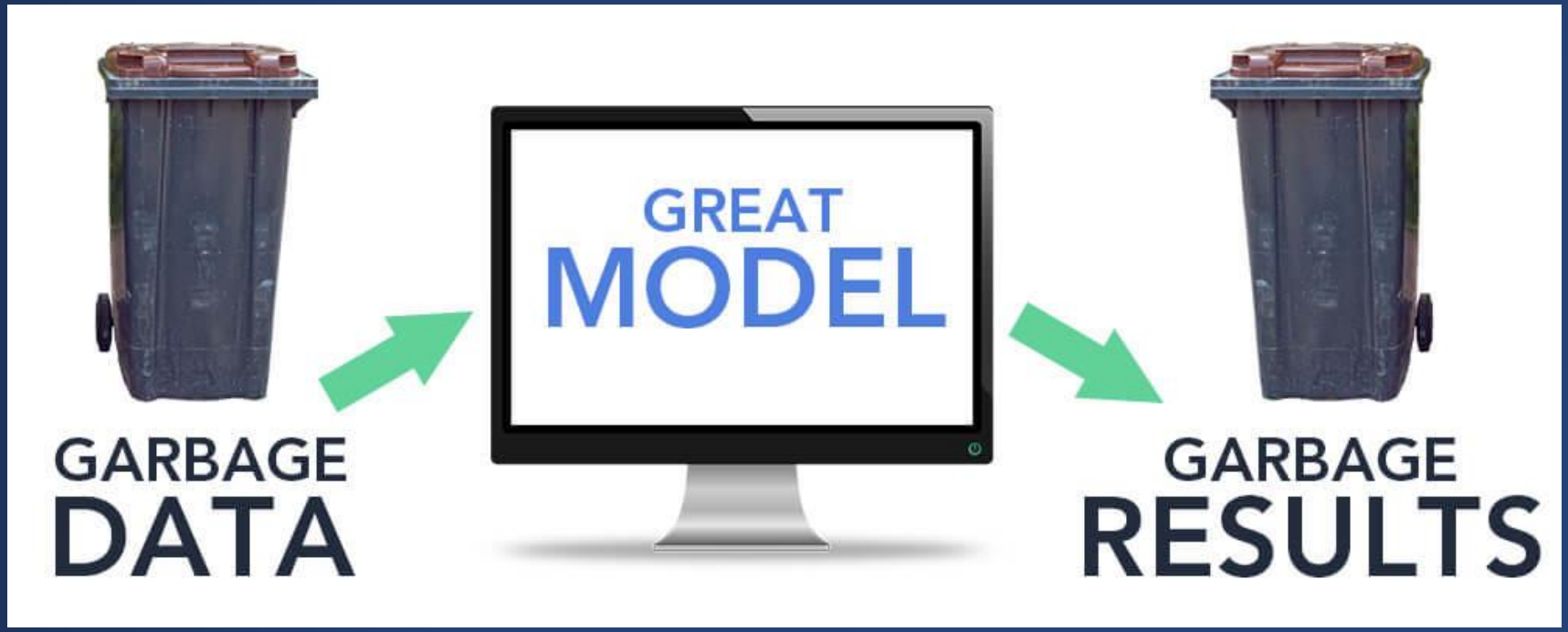
AI in Specific Industries

- AI subject to laws/regs of industry
  - Healthcare
  - Automobiles
  - Defense
  - Energy
  - Manufacturing

Data Issues

- "Biased" data
  - Too little, too skewed – over/under-sampling – proxy
  - Human/user-generated – "real world"
  - Labeled – culturally calibrated and inclusive
  - Confidence score
- What's "fair"?
  - Group, individual, process, results – balance

Data Issues



How Define?

"Fairness"

“21 Definitions of Algorithmic Fairness”

- There are more than 30 different mathematical definitions of fairness in the academic literature.
- There isn’t a one, true definition of fairness.
- These definitions can be grouped together into three families:
  - Anti-Classification
  - Classification Parity
  - Calibration

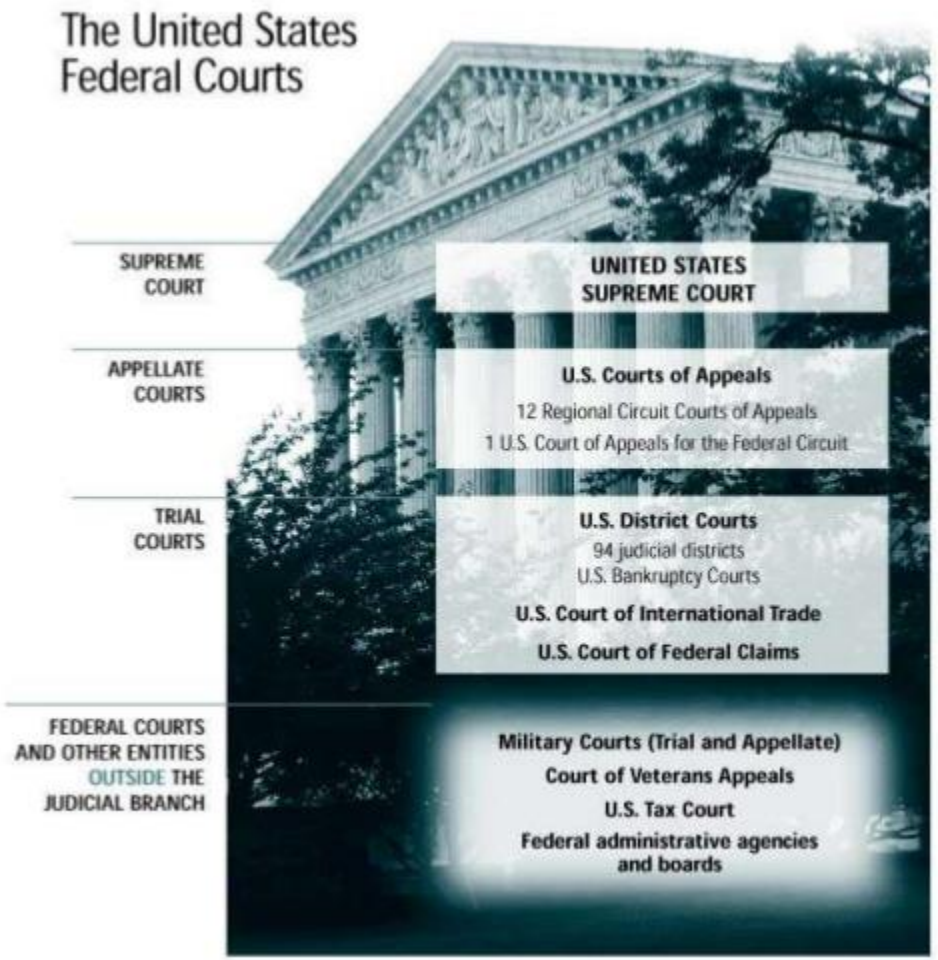
Arvind Narayanan

Individual Fairness

Group Fairness

Procedural Fairness

The US Federal Court System



Compositional Fairness

Counterfactual Fairness

If I had left the event early,

I would not have met

my soulmate.

Counterfactual

Outcome Fairness

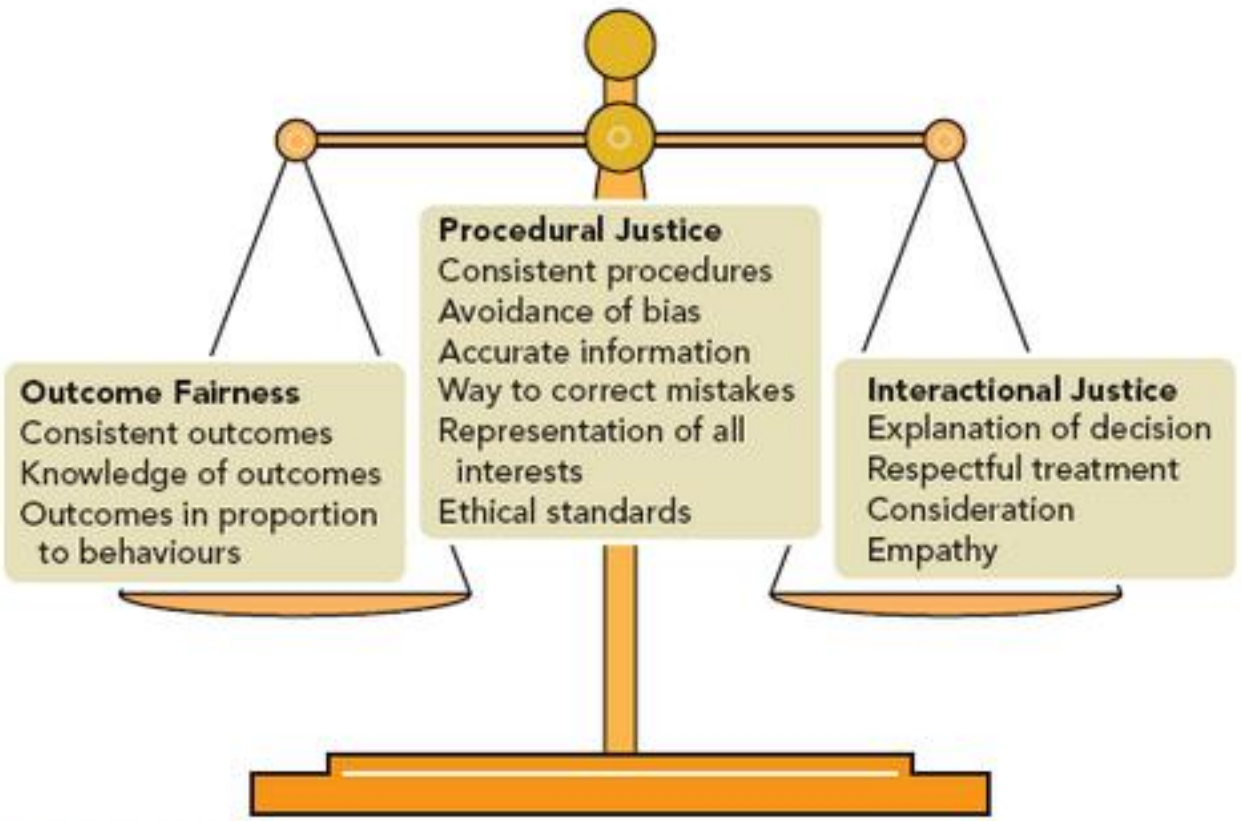


FIGURE 11.8

Principles of Justice

Combinations & tradeoffs

- Use multiple definitions?
- Tradeoffs?

Fairness & Accuracy

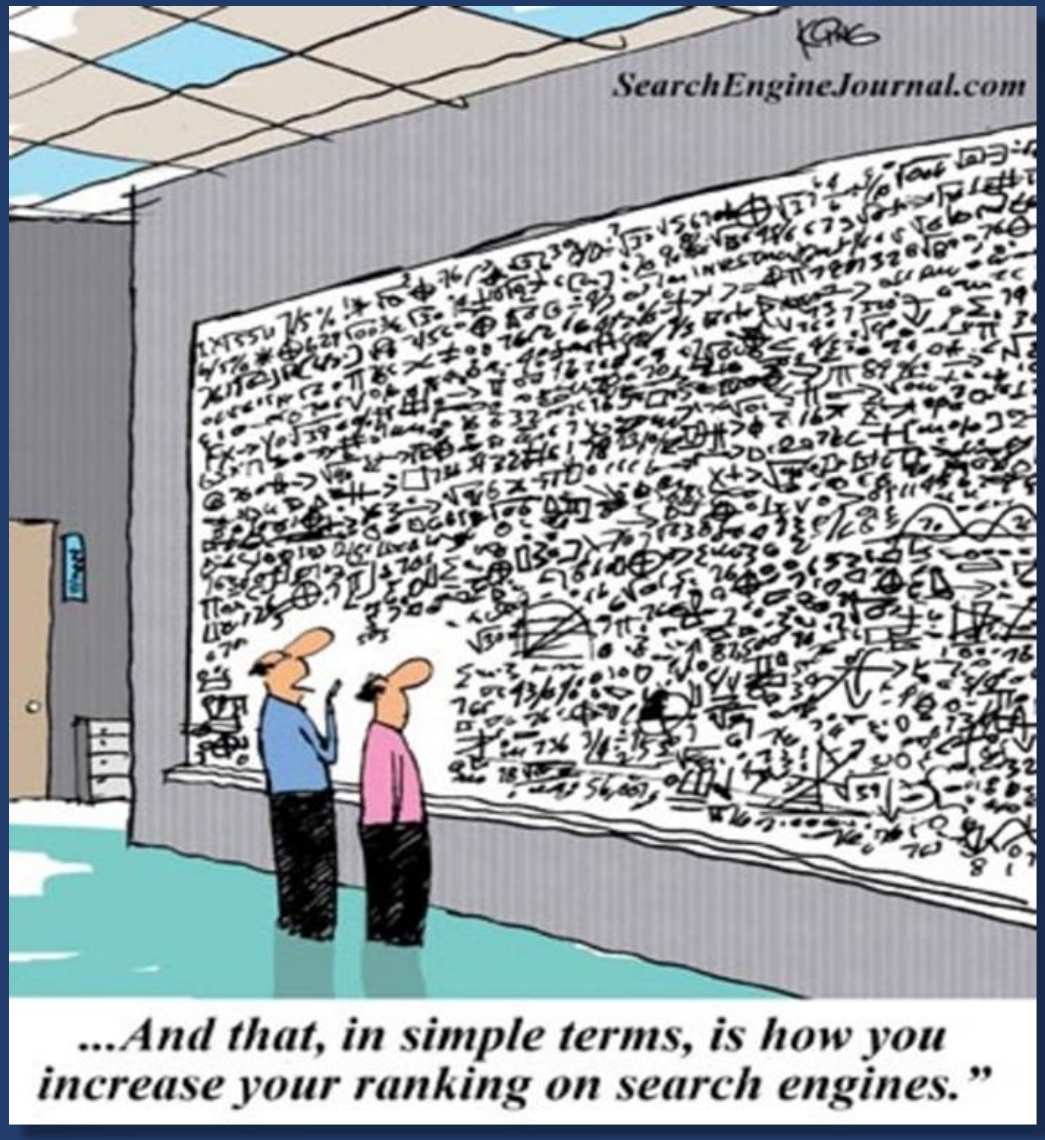
- Fairness affects accuracy
- Tradeoff Accuracy & Fairness
- Fairness constraints backfire?
  - Loans to those who cannot pay them back can negatively affect their credit score

Explainability

- If AI denies:
  - Loan
  - Bail/parole
  - Promotion or job
- Explanation?

EXPLAINABILITY

EXPLAINABILITY



Explainability

EXPLAINABILITY

Testing/validation – *e.g.*, drug trials

Explainability

If I had left the event early,

I would not have met

my soulmate.

Counterfactual

Explainability

- "If your income was $10,000/year more, then your loan would have been approved"
- Explanation consistent with fair process

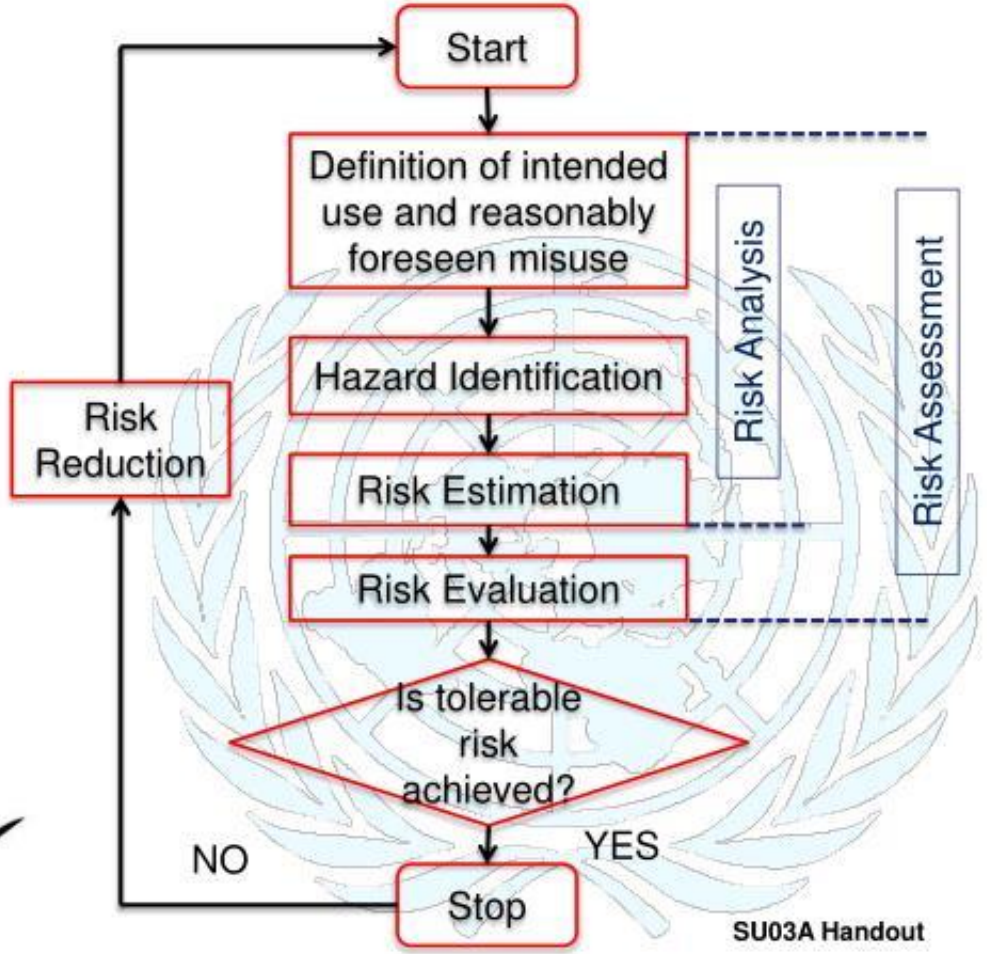
Recommendations

- Treat data and algorithmic bias as defects that must be investigated and overseen throughout the product development lifecycle and monitored post-sale

RECOMMENDATIONS

Risk Assessment Flow Chart

Remember – we are trying to reduce risk to tolerable and ALARP



UN SaferGuard
Securing ammunition, protecting lives

SU03A Handout