Durham
University

COMP 3647
Human-AI Interaction Design

Topic 15
Ethics in AI

Dr. Swaroop Panda,
Prof. Effie Law

Roadmap

- Tech Ethics
- Explainable & Responsible AI
- ML Fairness
- Bias & Discrimination
- Human Autonomy & Privacy
- Governance & Legal Aspects

What is Ethics

- Aristotle’s Ethics
- Normative Ethics
  - Virtue Ethics
  - Deontology
  - Consequentialism

Tech Ethics

- The Trolley Problem

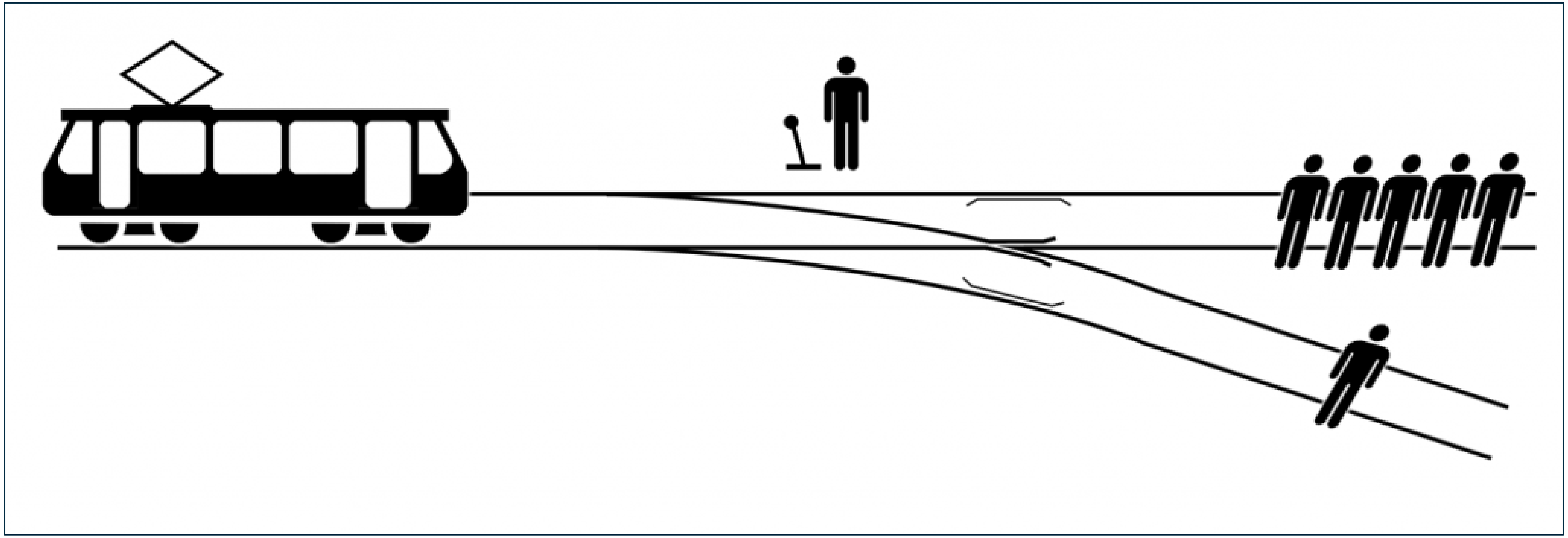


Image by McGeddon (CC BY-SA 4.0)

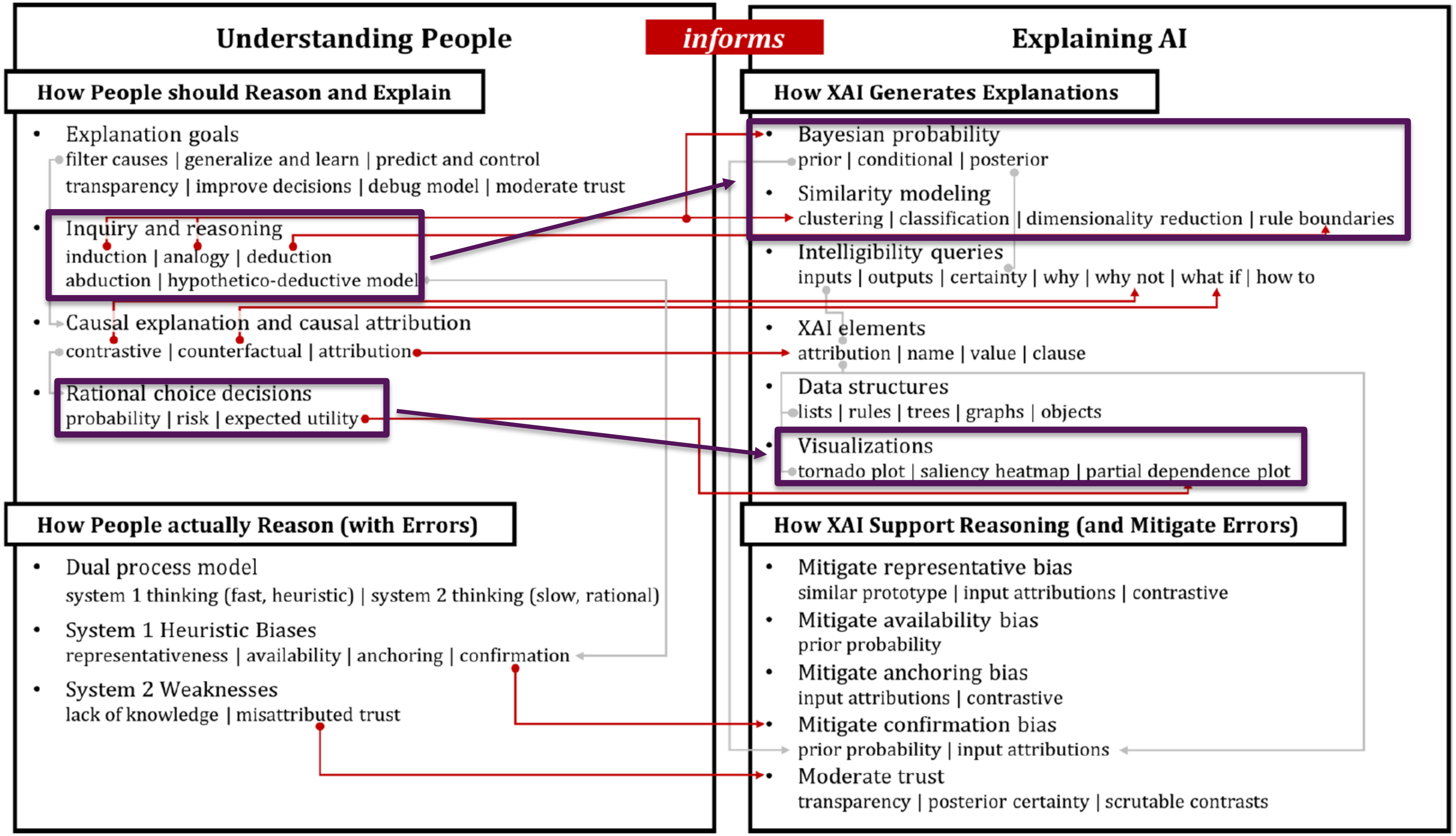
Tech Ethics

You have two options:

1. **Do nothing,** and the trolley kills the five people on the main track.
2. **Pull the lever,** diverting the trolley onto the side track where it will kill one person.

Which is the most **ethical choice?**

Explainable AI



Understanding People

informs

How People should Reason and Explain

- Explanation goals
  filter causes | generalize and learn | predict and control
  transparency | improve decisions | debug model | moderate trust
- Inquiry and reasoning
  induction | analogy | deduction
  abduction | hypothetico-deductive model
- Causal explanation and causal attribution
  contrastive | counterfactual | attribution
- Rational choice decisions
  probability | risk | expected utility

How People actually Reason (with Errors)

- Dual process model
  system 1 thinking (fast, heuristic) | system 2 thinking (slow, rational)
- System 1 Heuristic Biases
  representativeness | availability | anchoring | confirmation
- System 2 Weaknesses
  lack of knowledge | misattributed trust

Explaining AI

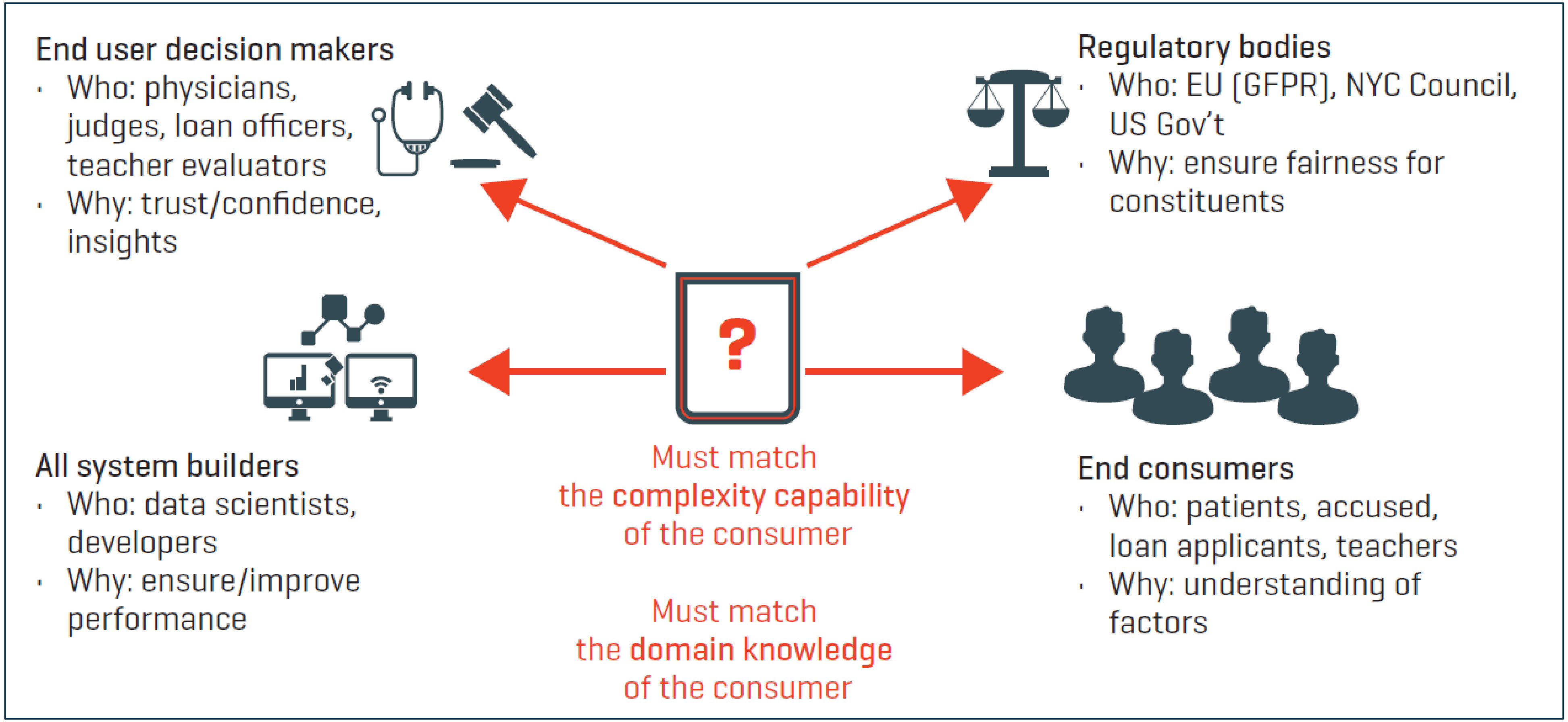
How XAI Generates Explanations

- Bayesian probability
  prior | conditional | posterior
- Similarity modeling
  clustering | classification | dimensionality reduction | rule boundaries
- Intelligibility queries
  inputs | outputs | certainty | why | why not | what if | how to
- XAI elements
  attribution | name | value | clause
- Data structures
  lists | rules | trees | graphs | objects
- Visualizations
  tornado plot | saliency heatmap | partial dependence plot

How XAI Support Reasoning (and Mitigate Errors)

- Mitigate representative bias
  similar prototype | input attributions | contrastive
- Mitigate availability bias
  prior probability
- Mitigate anchoring bias
  input attributions | contrastive
- Mitigate confirmation bias
  prior probability | input attributions
- Moderate trust
  transparency | posterior certainty | scrutable contrasts

Explainable AI



End user decision makers

- Who: physicians, judges, loan officers, teacher evaluators
- Why: trust/confidence, insights

Regulatory bodies

- Who: EU [GFPR], NYC Council, US Gov't
- Why: ensure fairness for constituents

All system builders

- Who: data scientists, developers
- Why: ensure/improve performance

End consumers

- Who: patients, accused, loan applicants, teachers
- Why: understanding of factors

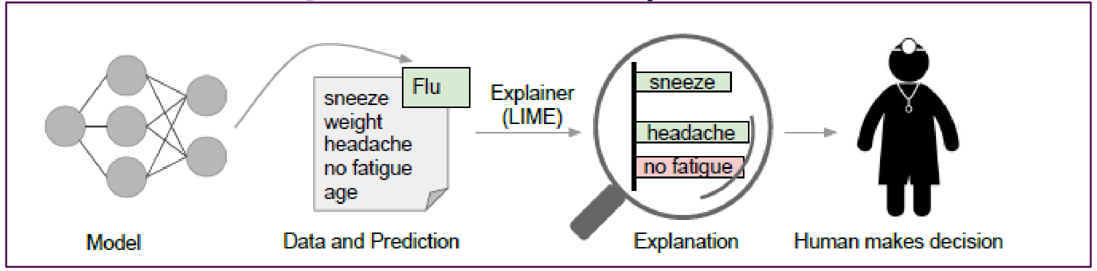
Must match
the **complexity capability**
of the consumer

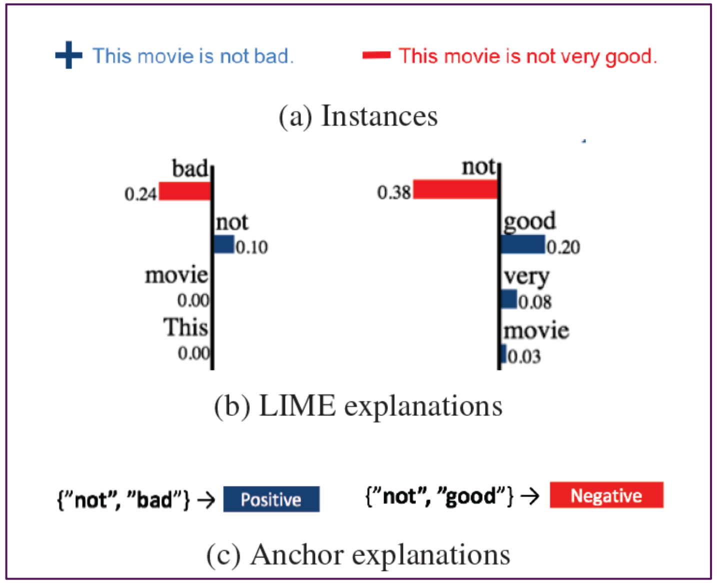
Must match
the **domain knowledge**
of the consumer

Explainable AI

Some XAI Methods,

1. LIME (Local Interpretable Model Agnostic Explanations)
2. Anchors
3. Layer-wise Relevance Propagation
4. Deep Taylor Decomposition (DTD)
5. Others



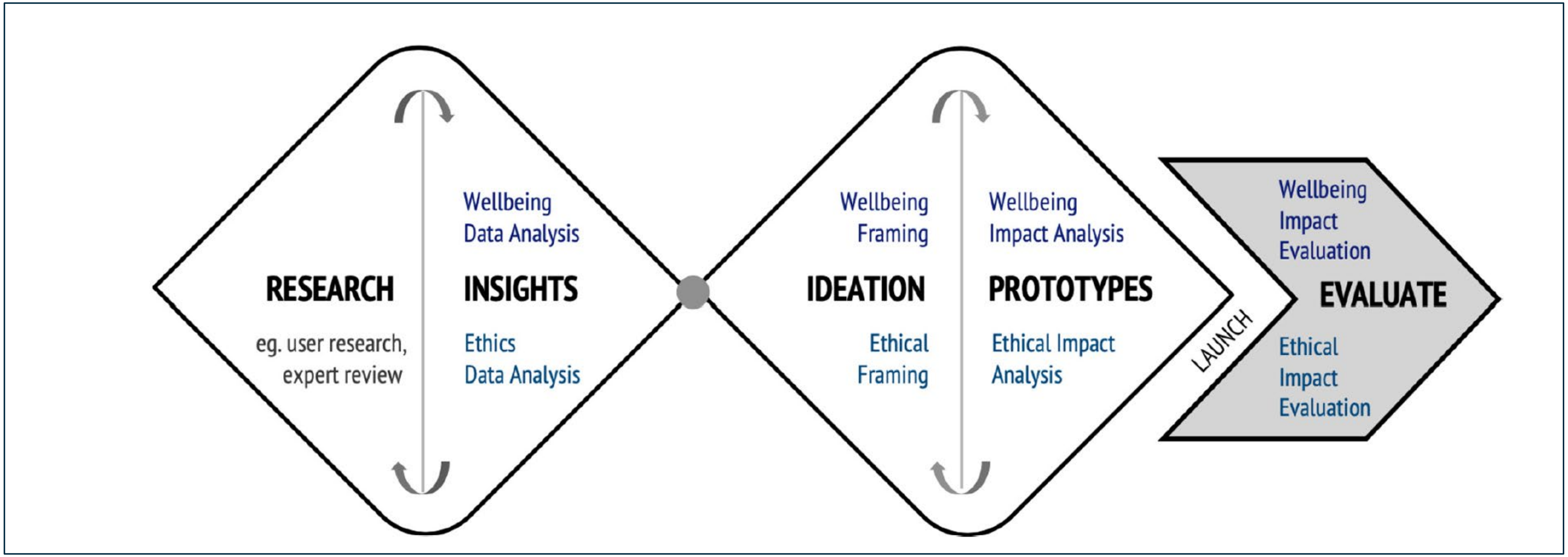


(a) Instances

(b) LIME explanations

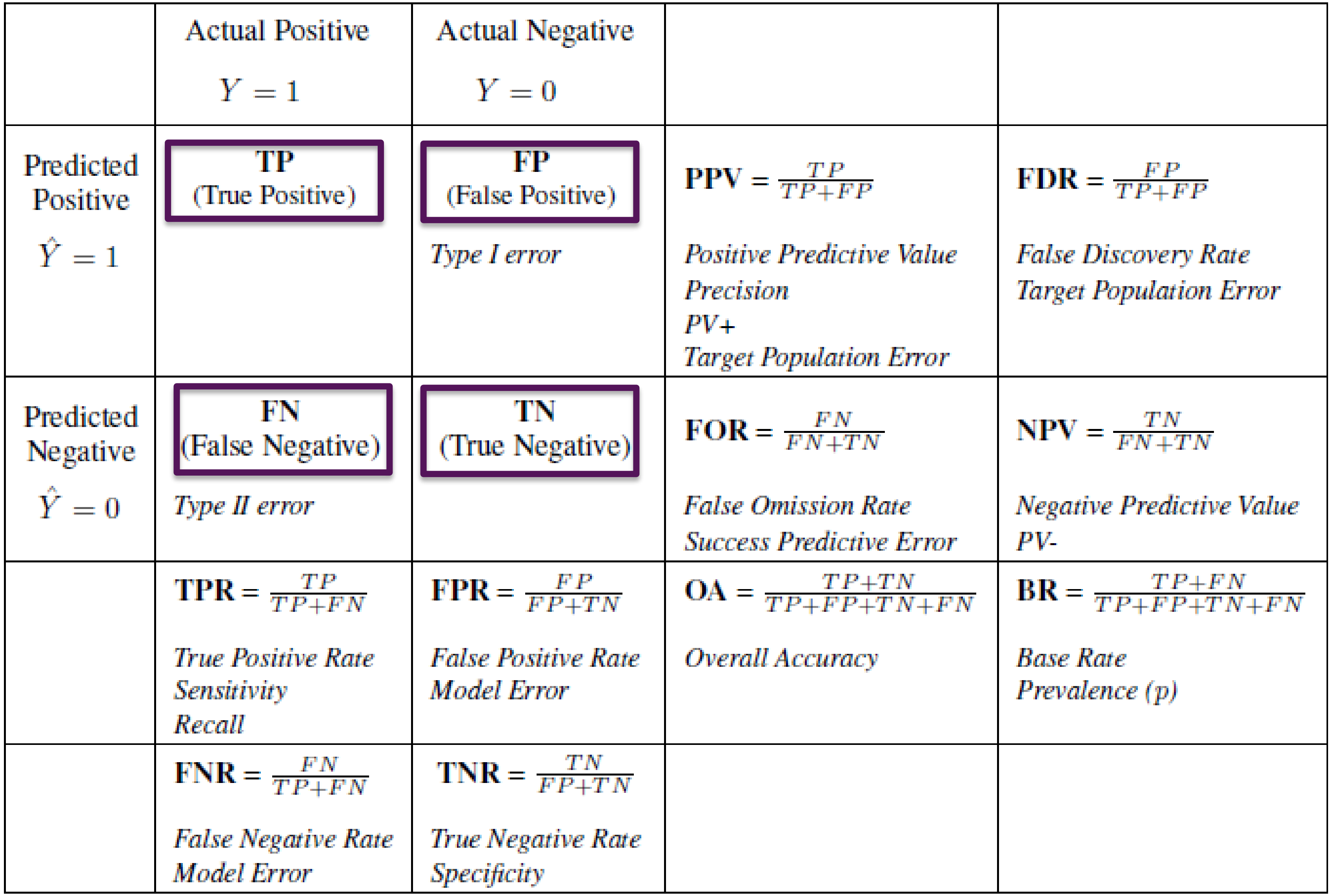
(c) Anchor explanations

Responsible AI



ML Fairness

Confusion Matrix



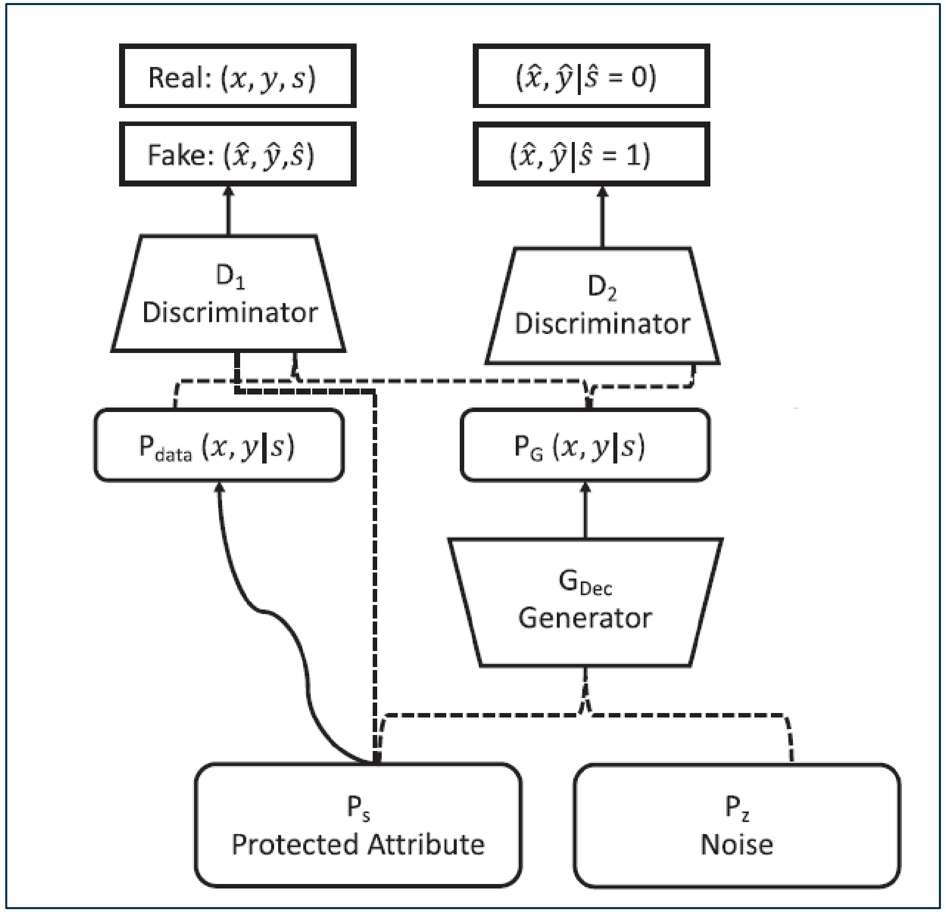
	Actual Positive $Y = 1$	Actual Negative $Y = 0$		
Predicted Positive $\hat{Y} = 1$	**TP** (True Positive)	**FP** (False Positive) *Type I error*	**PPV** = $\frac{TP}{TP+FP}$ *Positive Predictive Value* *Precision* *PV+* *Target Population Error*	**FDR** = $\frac{FP}{TP+FP}$ *False Discovery Rate* *Target Population Error*
Predicted Negative $\hat{Y} = 0$	**FN** (False Negative) *Type II error*	**TN** (True Negative)	**FOR** = $\frac{FN}{FN+TN}$ *False Omission Rate* *Success Predictive Error*	**NPV** = $\frac{TN}{FN+TN}$ *Negative Predictive Value* *PV-*
	TPR = $\frac{TP}{TP+FN}$ *True Positive Rate* *Sensitivity* *Recall*	**FPR** = $\frac{FP}{FP+TN}$ *False Positive Rate* *Model Error*	**OA** = $\frac{TP+TN}{TP+FP+TN+FN}$ *Overall Accuracy*	**BR** = $\frac{TP+FN}{TP+FP+TN+FN}$ *Base Rate* *Prevalence (p)*
	FNR = $\frac{FN}{TP+FN}$ *False Negative Rate* *Model Error*	**TNR** = $\frac{TN}{FP+TN}$ *True Negative Rate* *Specificity*		

ML Fairness

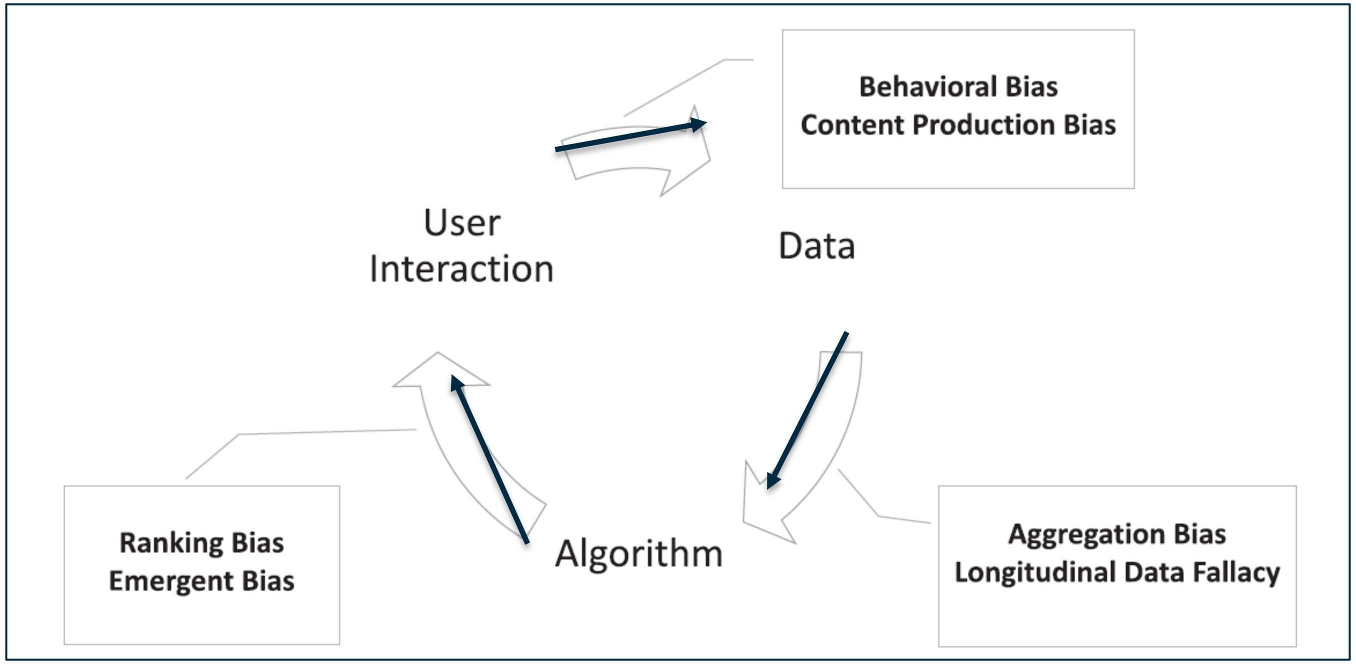
Definitions,

- Equalised odds
- Equal Opportunity
- Demographic Parity
- Fairness through awareness
- Test-fairness or calibration
- Others …

ML Fairness - FairGAN


Real: (x, y, s)
Fake: $(\hat{x}, \hat{y}, \hat{s})$
$(\hat{x}, \hat{y} | \hat{s} = 0)$
$(\hat{x}, \hat{y} | \hat{s} = 1)$
D_1 Discriminator
D_2 Discriminator
$P_{data}(x, y | s)$
$P_G(x, y | s)$
G_{Dec} Generator
P_s Protected Attribute
P_z Noise

ML Fairness


Behavioral Bias
Content Production Bias
User Interaction
Data
Ranking Bias
Emergent Bias
Algorithm
Aggregation Bias
Longitudinal Data Fallacy

Bias & Discrimination

Types of Bias

- Measurement Bias
- Omitted Variable Bias
- Representation Bias
- Aggregation Bias
  - Simpson’s Paradox.
  - Modifiable Areal Unit Problem

Bias & Discrimination

Types of Bias

- Algorithmic Bias
- User Interaction Bias
- Popularity Bias
- Emergent Bias
- Evaluation Bias
- Population Bias
- Historical Bias & Others

Bias & Discrimination

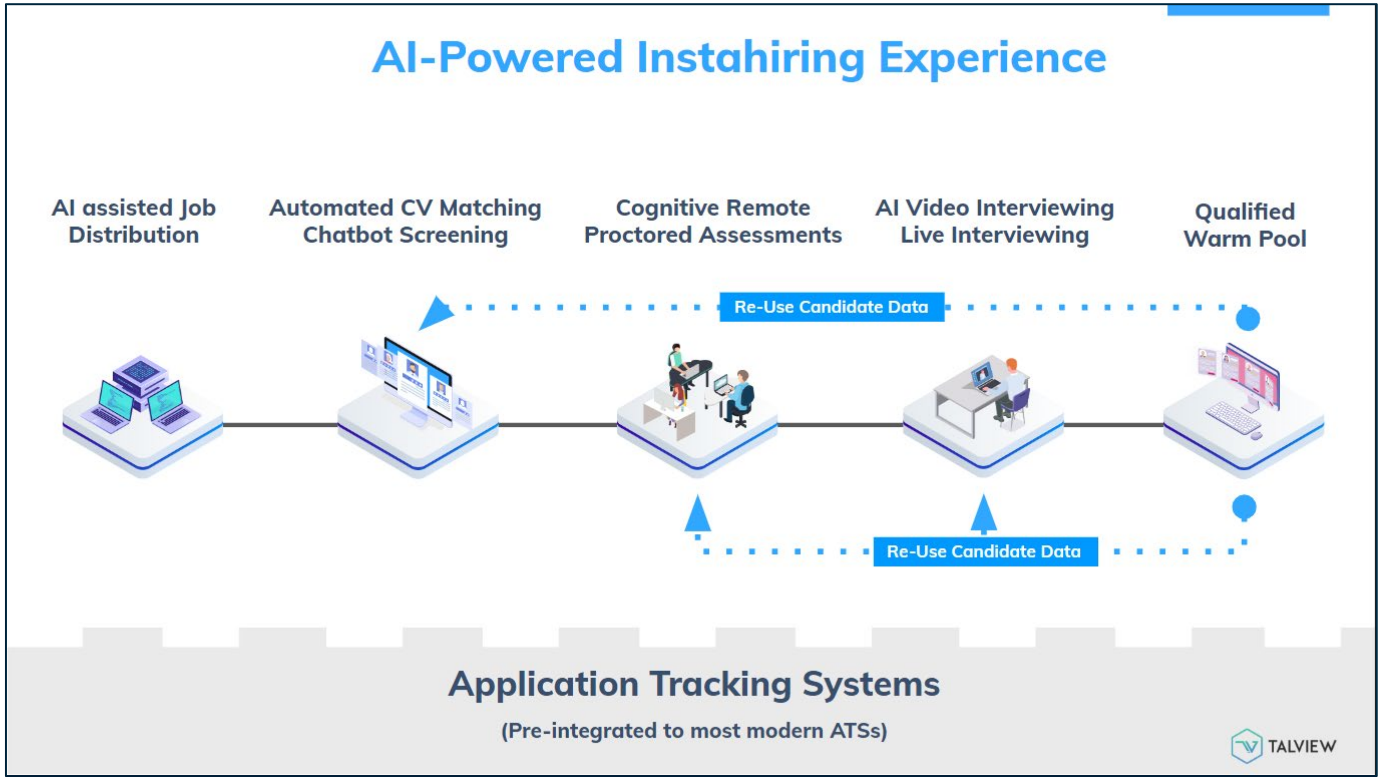
Discrimination vs **Bias**

Types of Discrimination,

- Systemic Discrimination
- Statistical Discrimination

Human Autonomy

- Replacing Human Labor



AI-Powered Instahiring Experience

AI assisted Job Distribution
Automated CV Matching Chatbot Screening
Cognitive Remote Proctored Assessments
AI Video Interviewing Live Interviewing
Qualified Warm Pool
Re-Use Candidate Data
Re-Use Candidate Data

Application Tracking Systems

(Pre-integrated to most modern ATSs)

Human Autonomy

- Autonomous Weapons

Human Autonomy

- Replacing Humans

Privacy

- Is Siri/Alexa hearing us?

Privacy

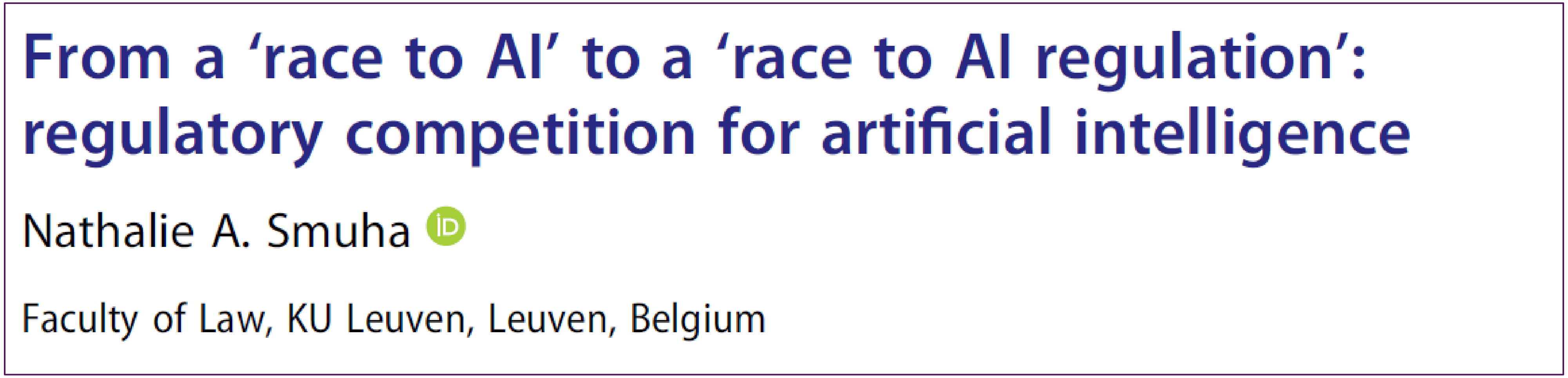
Surveillance

Privacy

- Facial Recognition

Legal Aspects

- Who should regulate AI?



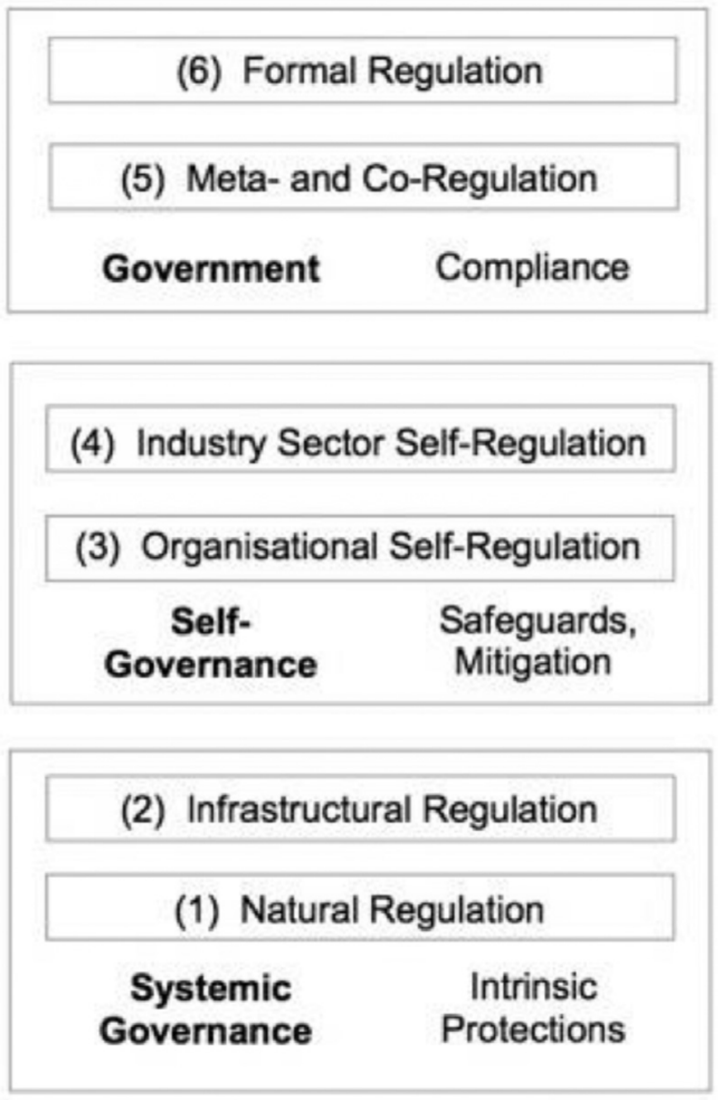
From a ‘race to AI’ to a ‘race to AI regulation’: regulatory competition for artificial intelligence

Nathalie A. Smuha

Faculty of Law, KU Leuven, Leuven, Belgium

Legal Aspects

- Who should regulate AI?



(6) Formal Regulation

(5) Meta- and Co-Regulation

Government Compliance

(4) Industry Sector Self-Regulation

(3) Organisational Self-Regulation

Self-Governance Safeguards, Mitigation

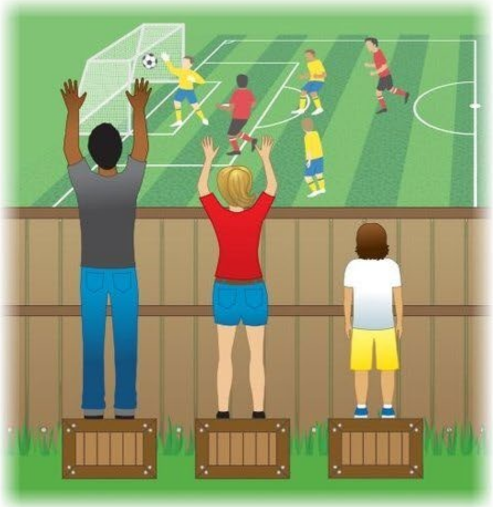
(2) Infrastructural Regulation

(1) Natural Regulation

Systemic Governance Intrinsic Protections

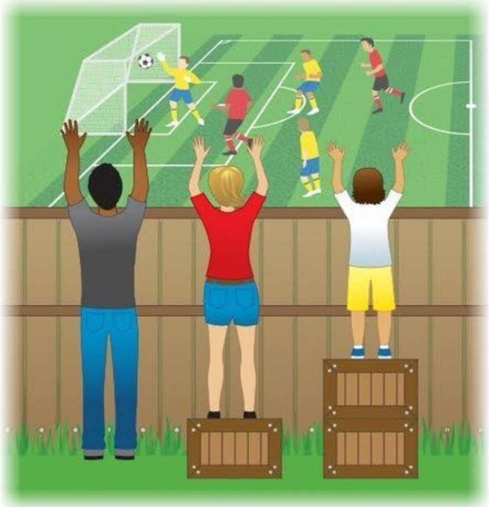
Legal Aspects

Equality



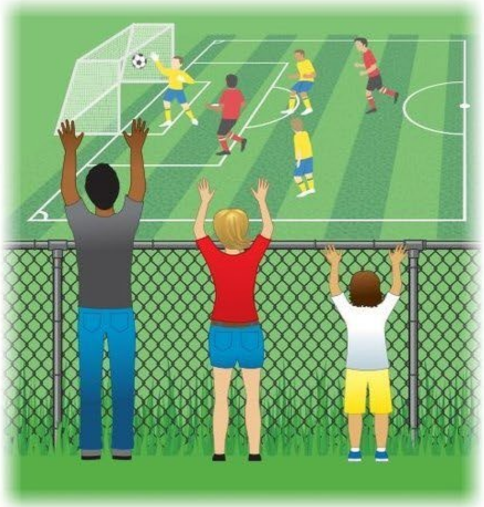
The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity



Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

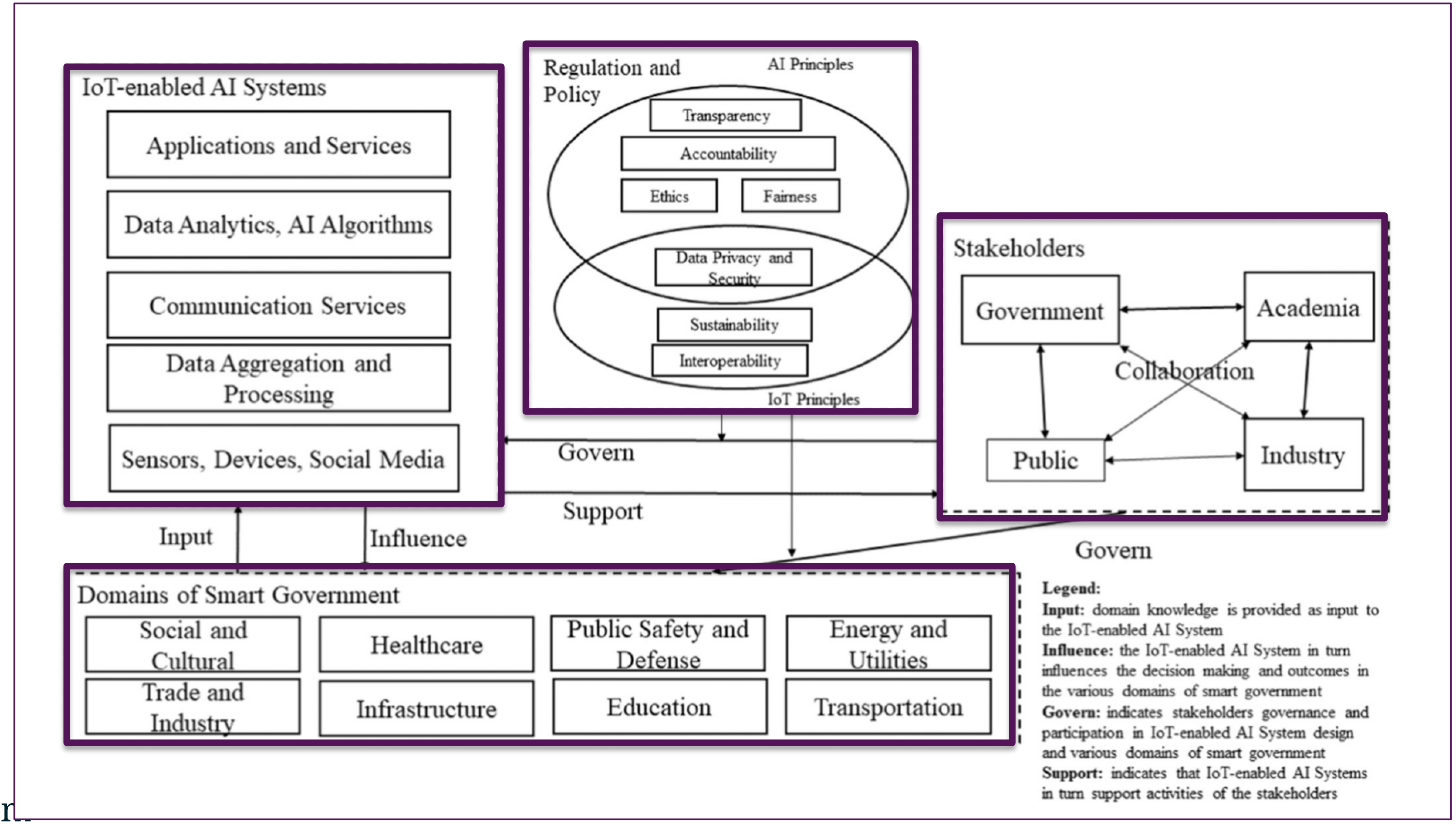
Justice



All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

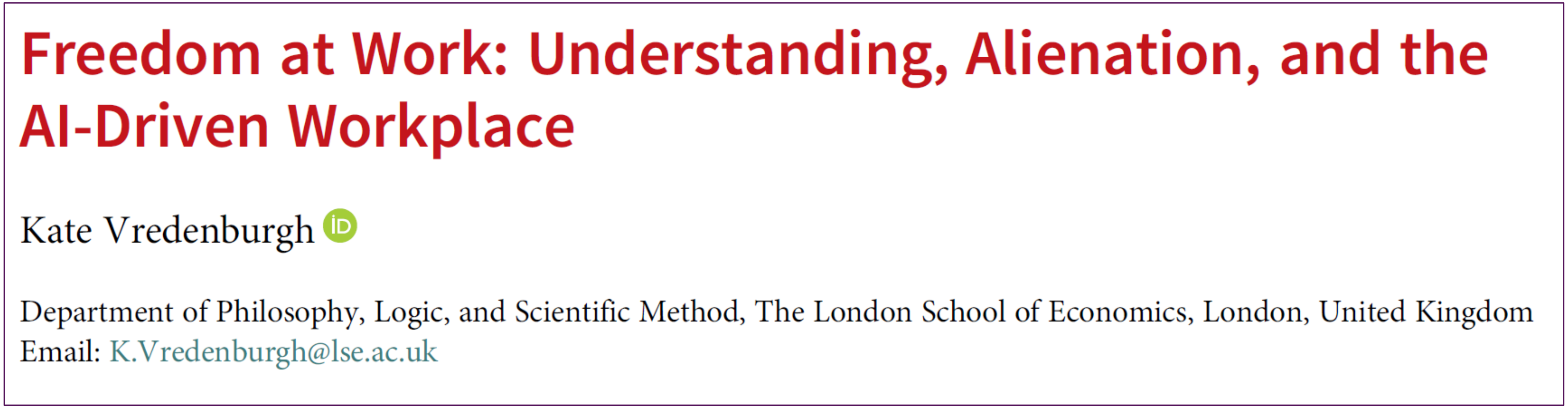
Governance

- Smart Governance



Legend:
Input: domain knowledge is provided as input to the IoT-enabled AI System
Influence: the IoT-enabled AI System in turn influences the decision making and outcomes in the various domains of smart government
Govern: indicates stakeholders governance and participation in IoT-enabled AI System design and various domains of smart government
Support: indicates that IoT-enabled AI Systems in turn support activities of the stakeholders

Governance

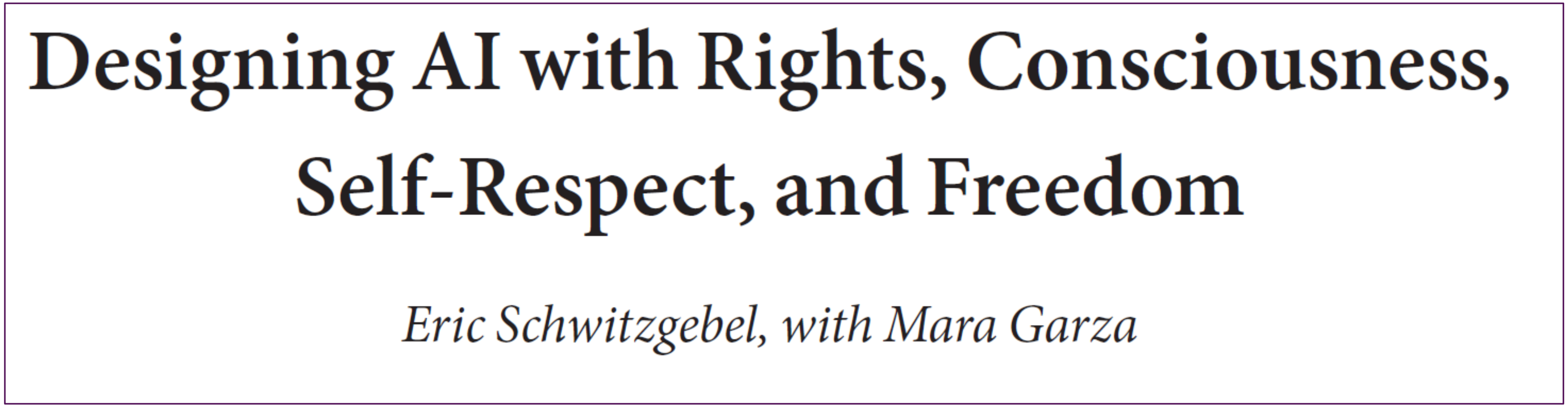


Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace

Kate Vredenburgh

Department of Philosophy, Logic, and Scientific Method, The London School of Economics, London, United Kingdom
Email: K.Vredenburgh@lse.ac.uk

Governance

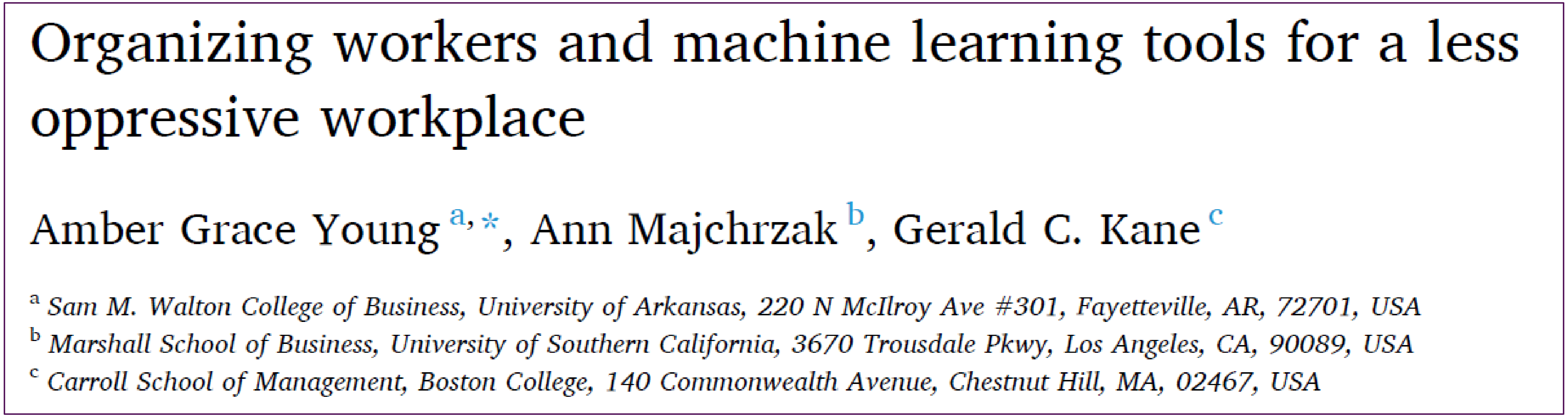


Designing AI with Rights, Consciousness, Self-Respect, and Freedom

Eric Schwitzgebel, with Mara Garza

Governance

- AI to oppress

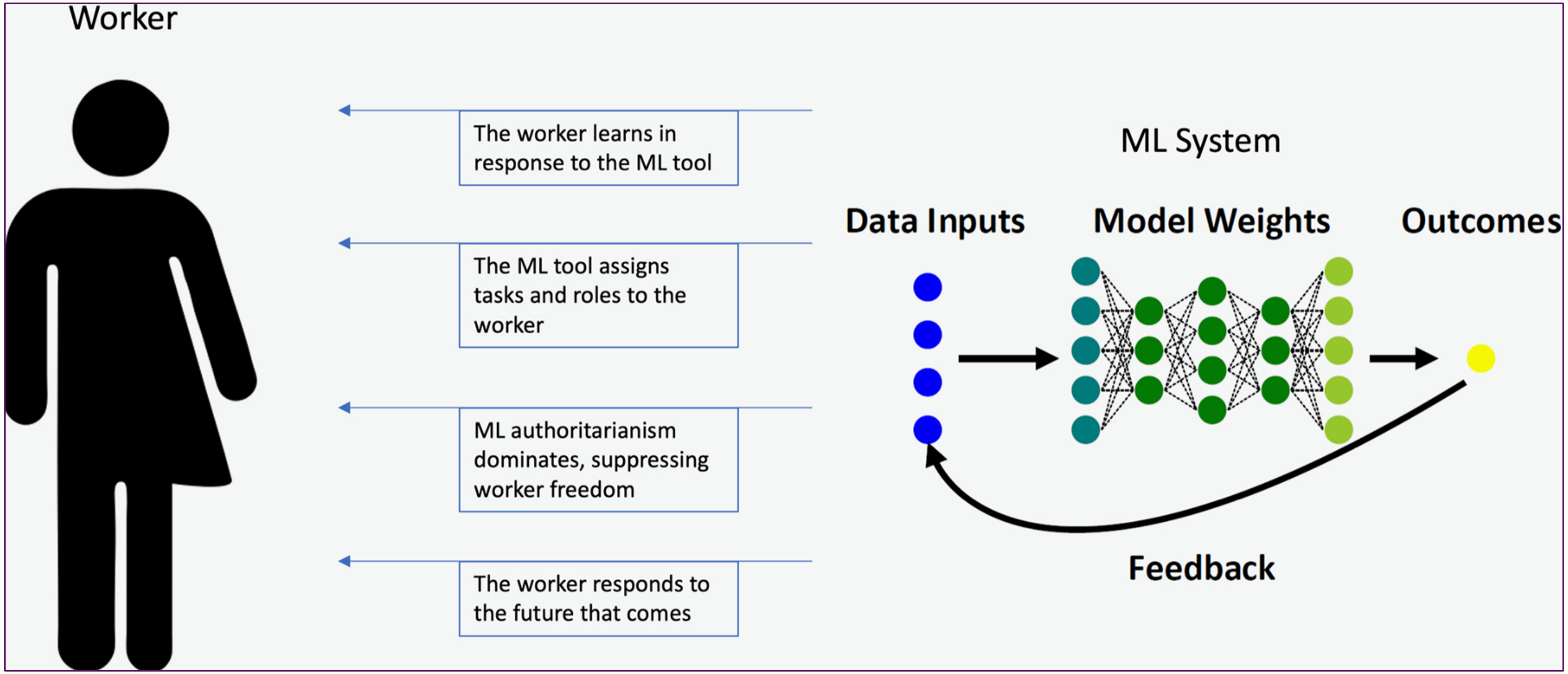


Organizing workers and machine learning tools for a less oppressive workplace

Amber Grace Young [a,*], Ann Majchrzak [b], Gerald C. Kane [c]

[a] Sam M. Walton College of Business, University of Arkansas, 220 N McIlroy Ave #301, Fayetteville, AR, 72701, USA
[b] Marshall School of Business, University of Southern California, 3670 Trousdale Pkwy, Los Angeles, CA, 90089, USA
[c] Carroll School of Management, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA, 02467, USA

Governance



Worker

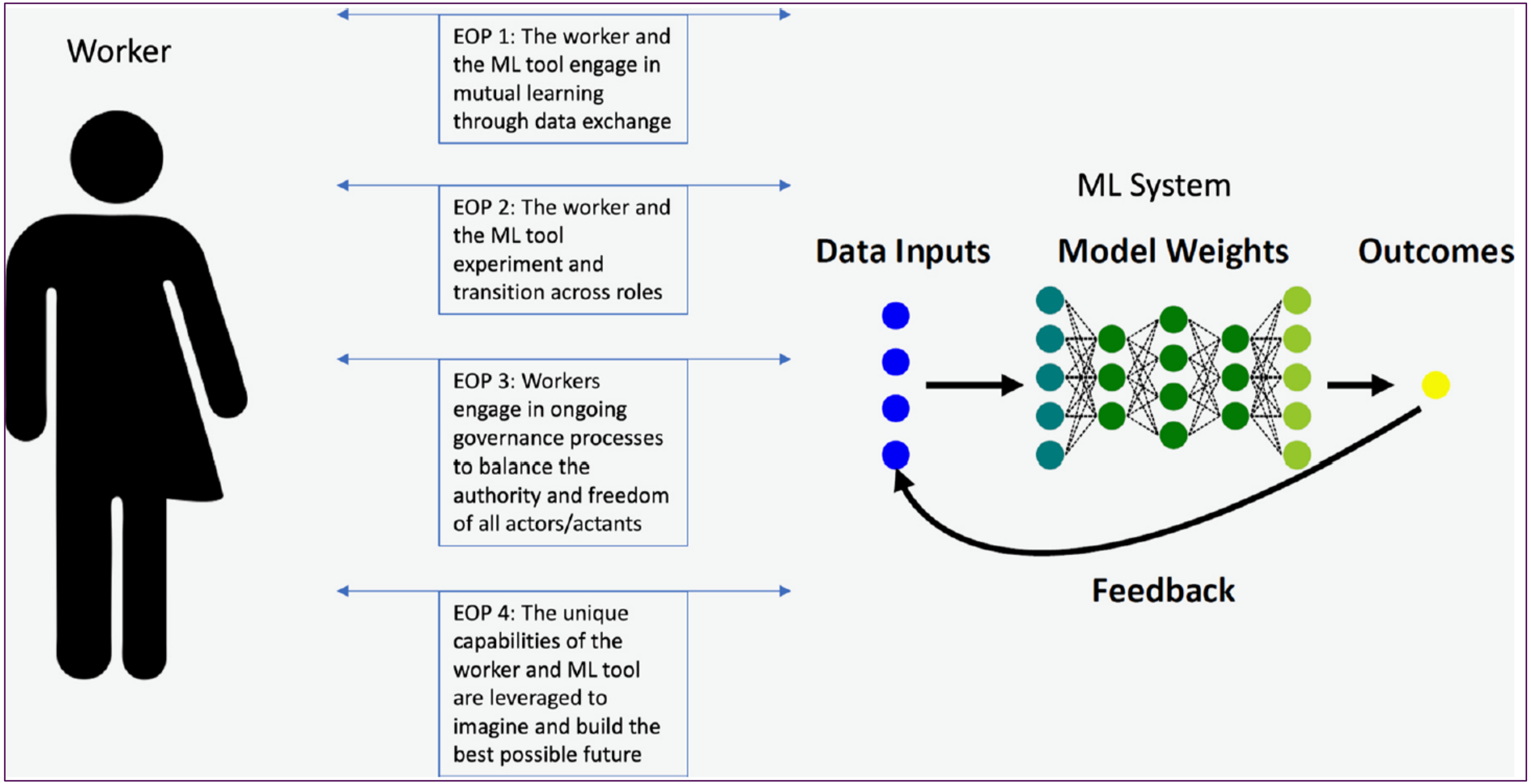
The worker learns in response to the ML tool

The ML tool assigns tasks and roles to the worker

ML authoritarianism dominates, suppressing worker freedom

The worker responds to the future that comes

Governance



EOP 1: The worker and the ML tool engage in mutual learning through data exchange

EOP 2: The worker and the ML tool experiment and transition across roles

EOP 3: Workers engage in ongoing governance processes to balance the authority and freedom of all actors/actants

EOP 4: The unique capabilities of the worker and ML tool are leveraged to imagine and build the best possible future

Wrapping Up

- Ethics in Technology
- Explainable & Responsible AI Methods
- Fairness in AI
- Bias & Discrimination
- Human Autonomy & Privacy
- Governance & Legal Aspects

Thanks

Any Questions?