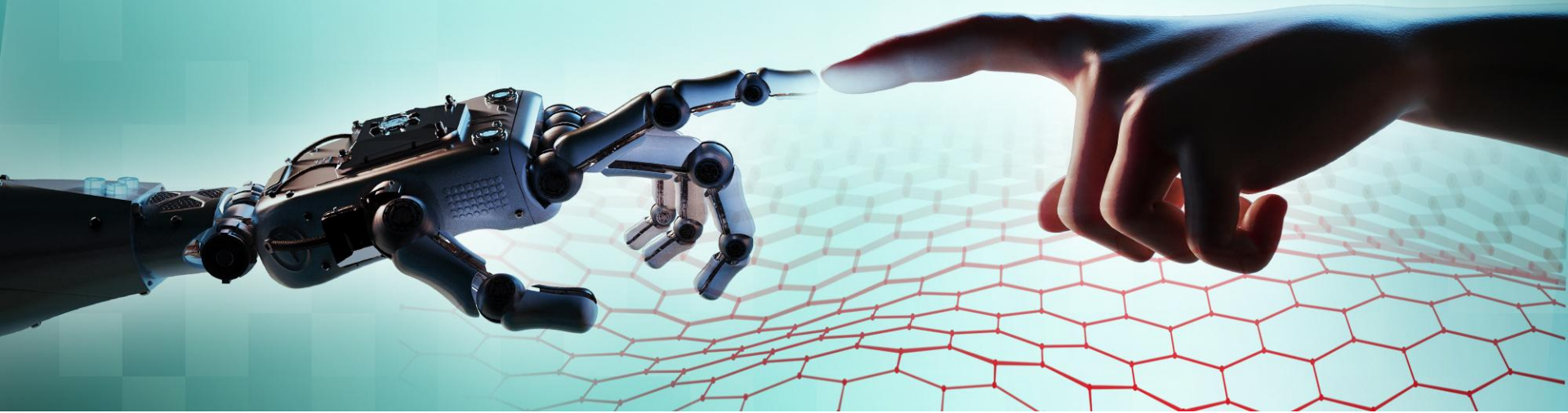
Governing the Societal Impacts of Generative AI:
Toward Better Evaluations



Hoda Heidari
K&L Gates Career Development
Assistant Professor in Ethics and
Computational Technologies

K&L Gates Initiative in
ETHICS & COMPUTATIONAL
TECHNOLOGIES

Carnegie Mellon University
Responsible AI

In-class Poll

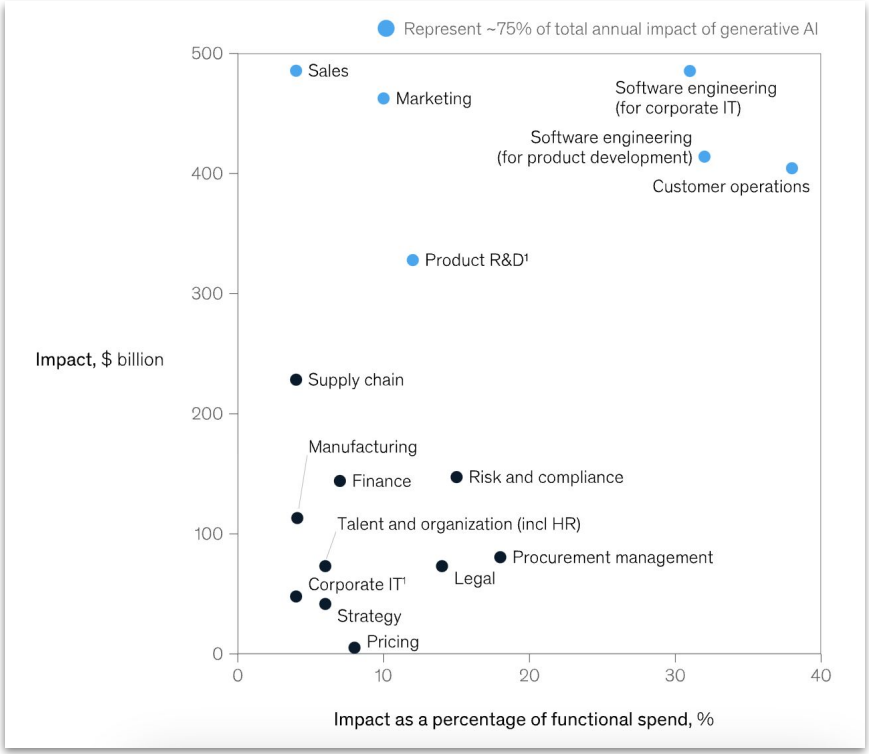
1. Which capability/societal benefit of GenAI are you most excited about?
2. Which risk/negative consequence of GenAI are you most worried about?
3. Should AI be regulated?

Outline

1. **Responsible Governance of AI**
    - Why? Who? How?

2. **GenAI Risk Evaluations**
    - Layer 1: Model testing
    - Layer 2: Red-teaming
    - Layer 3: Field testing

AI in Society: Benefits

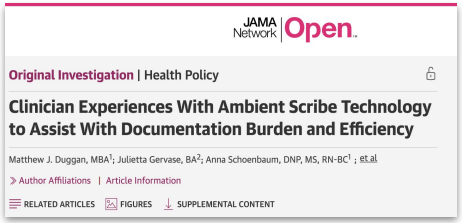


Source: McKinsey Report on "The economic potential of generative AI: The next productivity frontier"

January 24, 2025 ACCESS TO JUSTICE

Access to Justice 2.0: How AI-powered software can bridge the gap

By Nicole Black



JAMA Network | Open.

Original Investigation | Health Policy

Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency

Matthew J. Duggan, MBA[1]; Julietta Gervase, BA[2]; Anna Schoenbaum, DNP, MS, RN-BC[1] ; et al

Author Affiliations | Article Information

RELATED ARTICLES FIGURES SUPPLEMENTAL CONTENT

nature › scientific reports › articles › article

Article | Open access | Published: 03 June 2025

AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting

Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne & Gregorio Ponti

Scientific Reports 15, Article number: 17458 (2025) | Cite this article

12k Accesses | 3 Citations | 67 Altmetric | Metrics

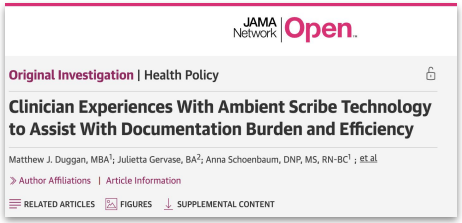
AI in Society: Benefits

- Innovation and economic growth
- Productivity and efficiency gains
- Due process
  - Consistency
  - Traceability
  - Making choices & biases evident
- …

January 24, 2025 ACCESS TO JUSTICE

Access to Justice 2.0: How AI-powered software can bridge the gap

By Nicole Black



JAMA Network | Open.

Original Investigation | Health Policy

Clinician Experiences With Ambient Scribe Technology to Assist With Documentation Burden and Efficiency

Matthew J. Duggan, MBA[1]; Julietta Gervase, BA[2]; Anna Schoenbaum, DNP, MS, RN-BC[1] ; et al

Author Affiliations | Article Information

RELATED ARTICLES FIGURES SUPPLEMENTAL CONTENT

nature › scientific reports › articles › article

Article | Open access | Published: 03 June 2025

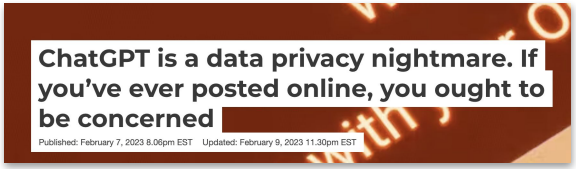
AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting

Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne & Gregorio Ponti

Scientific Reports **15**, Article number: 17458 (2025) | Cite this article

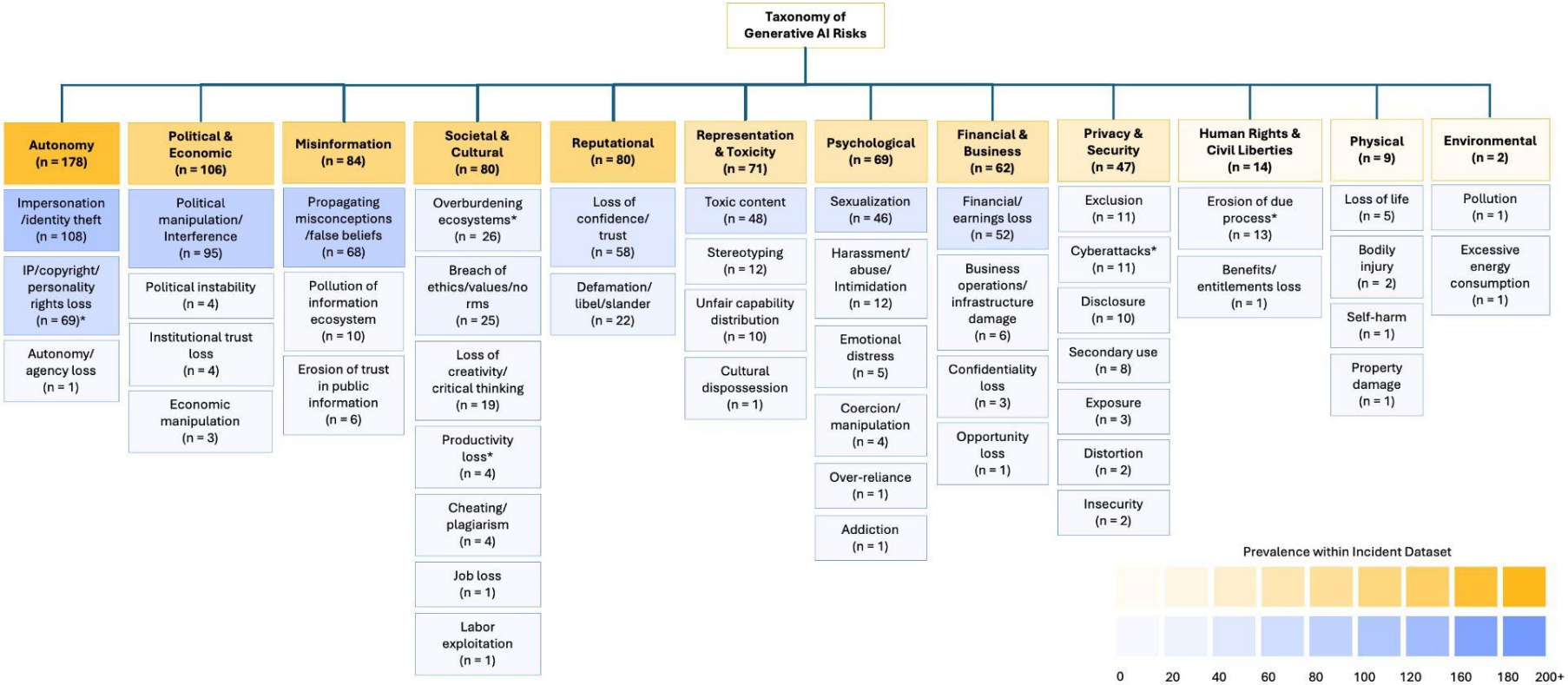
12k Accesses | **3** Citations | **67** Altmetric | Metrics

AI in Society: Harms



- Violations of human rights
  - Justice, equity, and non-discrimination
  - Privacy and non-surveillance
  - Freedom of communication and expression
  - Economic freedom
- Negative impact on human flourishing and wellbeing
  - Loss of human sovereignty and control
  - Human cognitive abilities
  - ...

AI in Society: Harms



* Figure from “A Closer Look at the Existing Risks of Generative AI:”| Li et al., AIES 2025

Responsible AI Goals

Foster:

- Innovation and economic growth
- Productivity and efficiency gains
- Due process
  - Consistency
  - Traceability
  - Making choices & biases evident
- ...

Mitigate:

- Violations of human rights
  - Justice, equity, and non-discrimination
  - Privacy and non-surveillance
  - Freedom of communication and expression
  - Economic freedom
- Negative impact on human flourishing and wellbeing
  - Loss of human sovereignty and control
  - Human cognitive abilities
- ...

* Slide from “Accountability & Governance for ML”. NeurIP 2024 Tutorial | H. Heidari & E. Black.

Definition of “Governance”

The **system** by which an establishment is **controlled**, and **directed** toward its goals and the **mechanisms** by which the establishment and its people, are held to **account**.

* Slide from “Accountability & Governance for ML”. NeurIP 2024 Tutorial | H. Heidari & E. Black.

How: Governance Challenges

1- Who should comply with what?

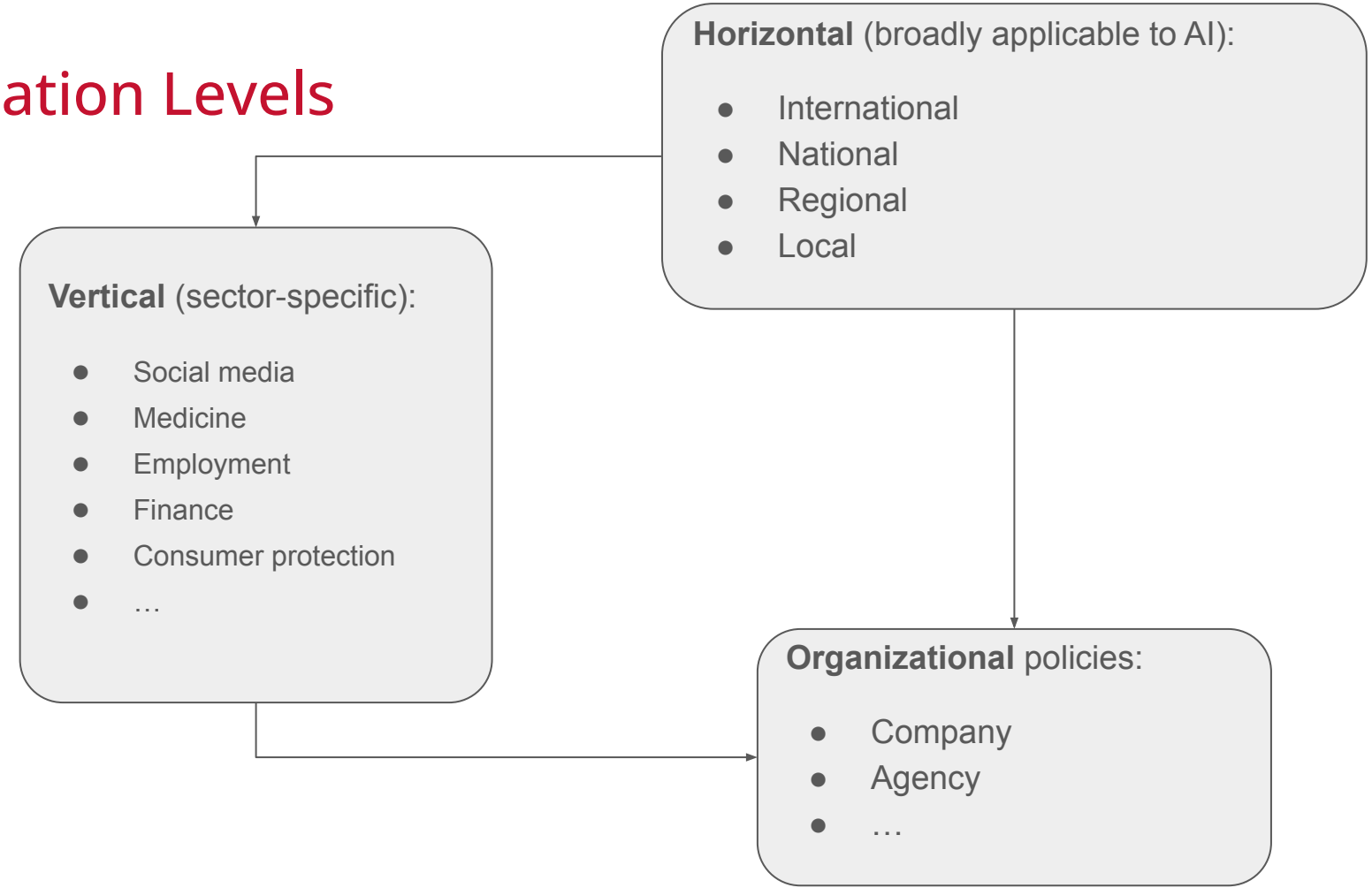
- **Multiple stakeholders**, goals, and values
  - tradeoffs among various stakeholders
  - accounting for different contexts
  - While avoiding conflicting policies
- **Lack of role-specific standards and best practices**
  - uncertain, fast-moving, dynamic environment
  - general-purpose technology
  - low rates of adoption and validation

2- How to incentivize compliance?

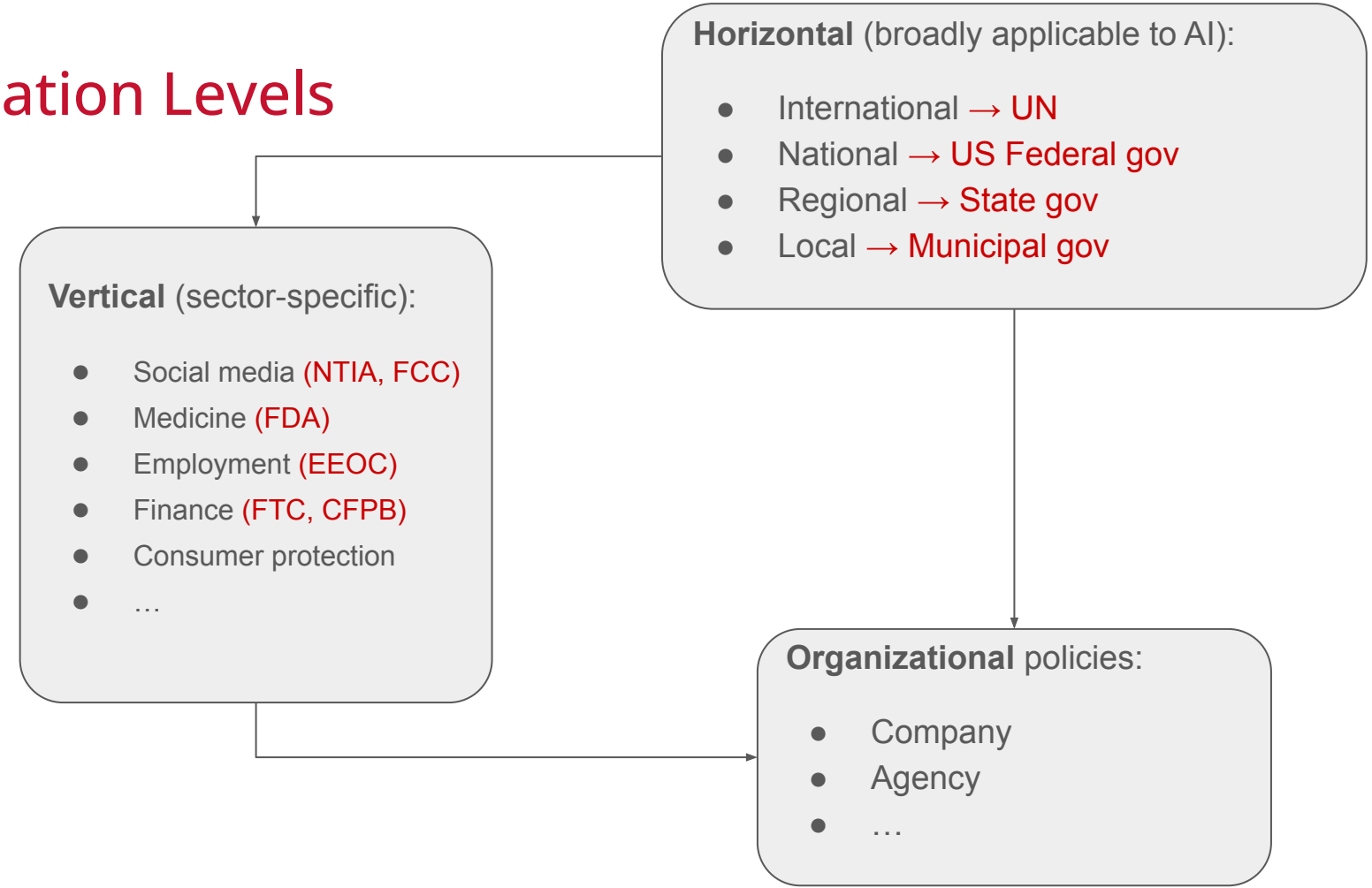
- **Assessing compliance**
  - Information asymmetries and access
  - Model opaqueness and inscrutability
- **Attributing liability** for harms
  - Harms are diffuse and hard to foresee
  - Intentionality, Recklessness, Negligence
  - creator/ owner/ operator/ user?
- **Enforcing consequences**
  - Balancing retributive justice and deterrence with innovation and growth

* Slide from “Accountability & Governance for ML”. NeurIP 2024 Tutorial | H. Heidari & E. Black.

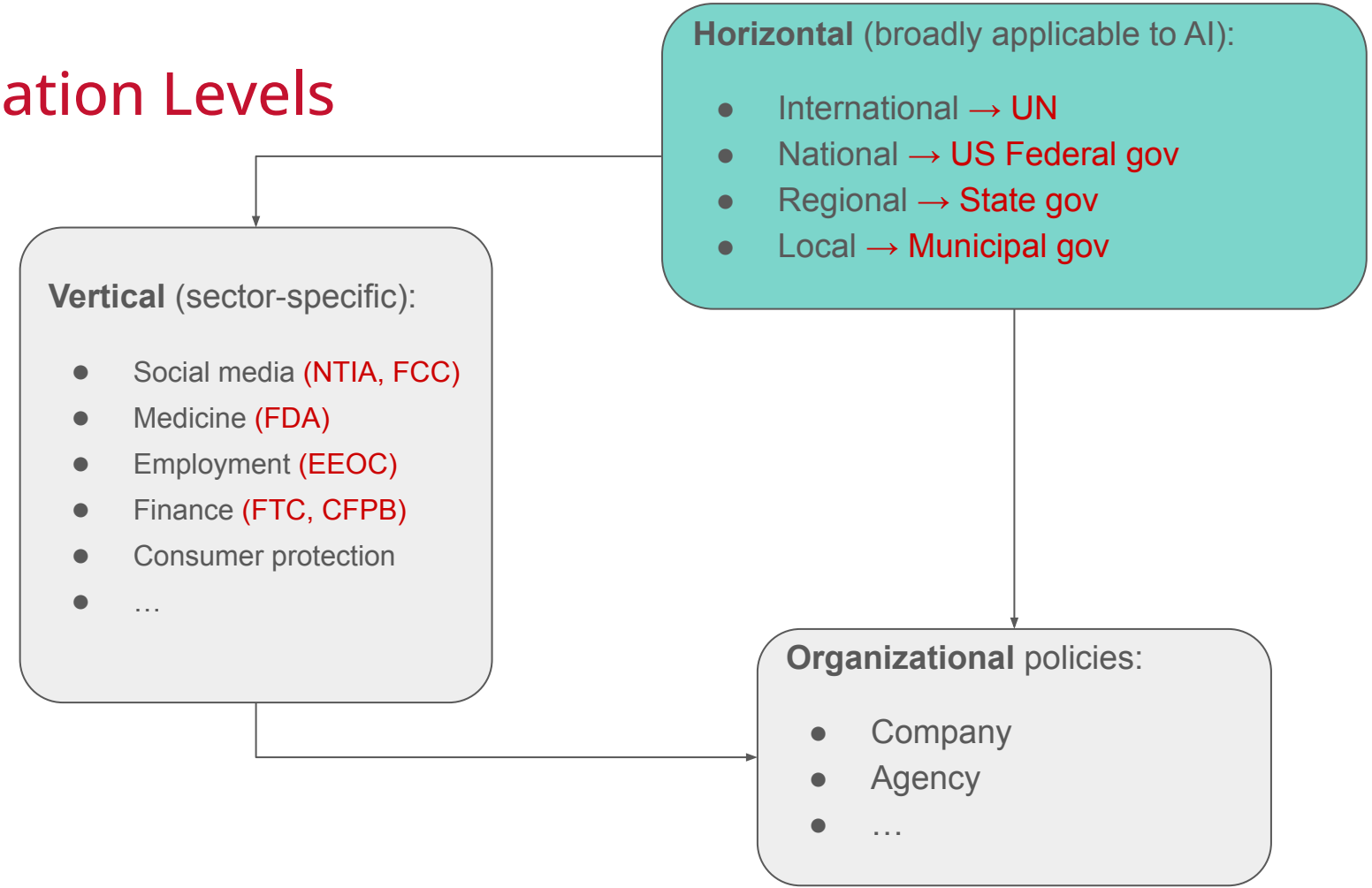
Regulation Levels



Regulation Levels

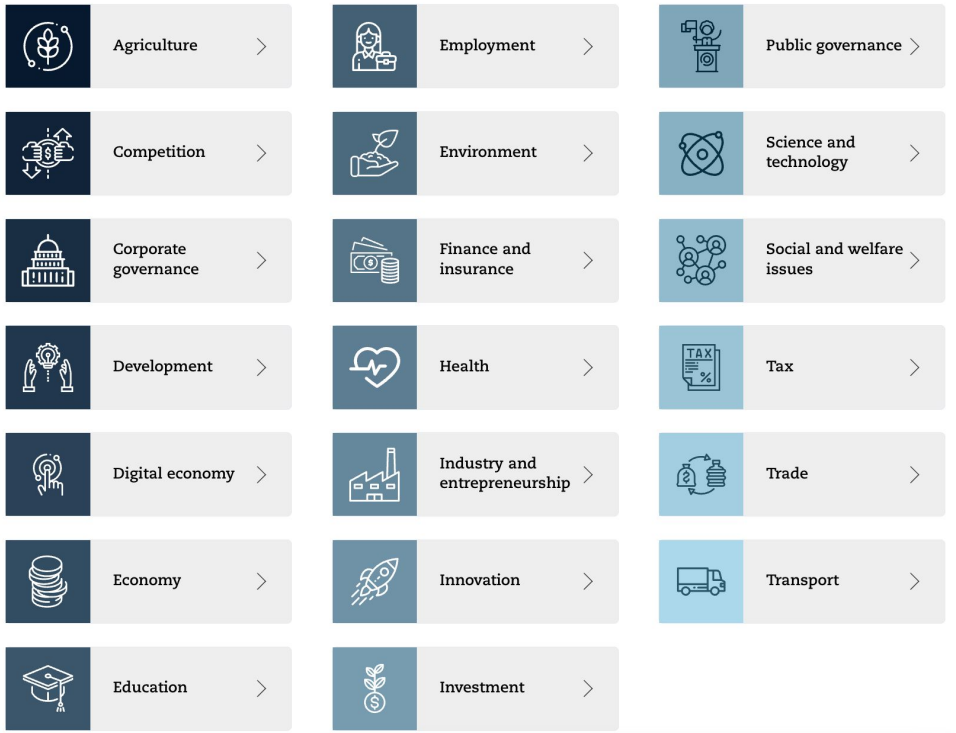


Regulation Levels

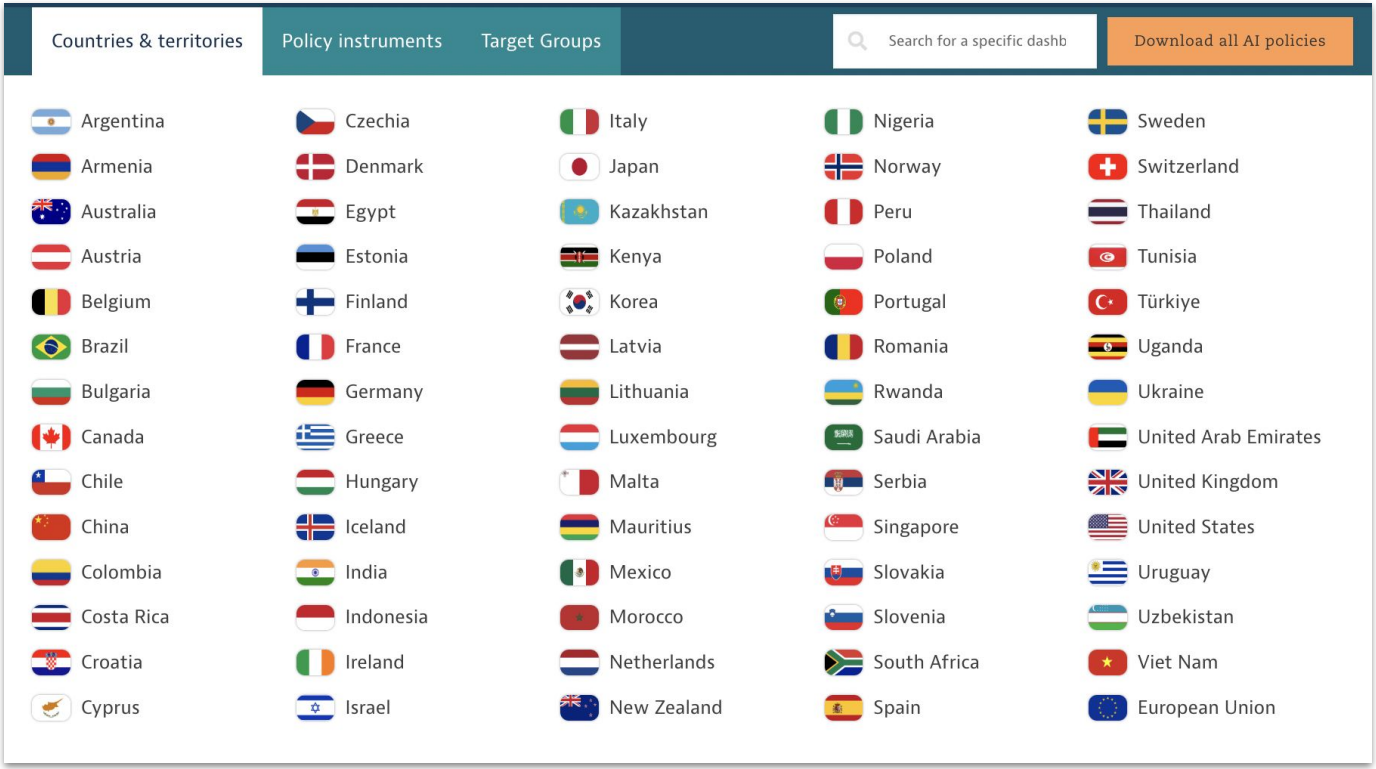


International

- OECD.AI
- UNESCO
- GPAI

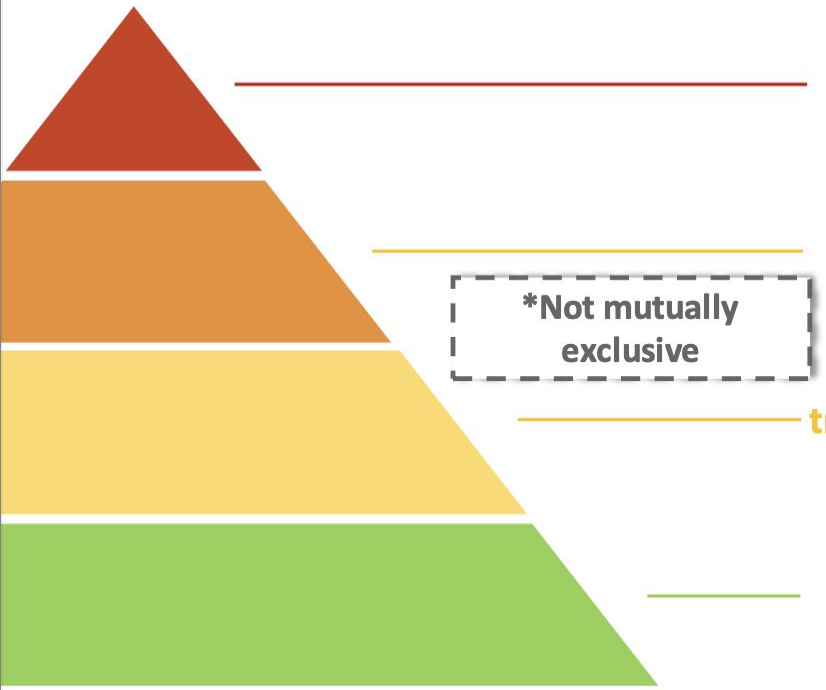


National AI Policy Tracker (OECD.AI)



https://oecd.ai/en/dashboards/national

The EU AI Act



*Not mutually exclusive

Unacceptable risk
e.g. social scoring

Prohibited

High risk
e.g. recruitment, medical devices

Permitted subject to compliance with AI requirements and ex-ante conformity assessment

AI with specific transparency obligations
'Impersonation' (bots)

Permitted but subject to information/transparency Obligations

Minimal or no risk

Permitted with no restrictions

European Commission

The EU AI Act - High-Risk Systems

1. Risk Management System
2. Data and Data Governance
3. Technical Documentation
4. Record-Keeping
5. Transparency and Information Disclosure
6. Human Oversight
7. Robustness, Accuracy, and Security
8. Conformity Assessment
9. Post-Market Monitoring
10. Registration in EU Database

https://artificialintelligenceact.eu/high-level-summary/

The EU AI Act - GPAI Systems

All providers must:

1. Draw up **technical documentation**, including training and testing process and evaluation results.
2. Draw up **information and documentation to supply to downstream providers** that intend to integrate the GPAI model into their own AI system in order that the latter understands capabilities and limitations and is enabled to comply.
3. Establish a policy to **respect the Copyright Directive**.
4. Publish a sufficiently **detailed summary about the content used for training** the GPAI model.

Providers of models with systemic risk:

1. Perform **model evaluations**, including conducting and documenting adversarial testing to identify and mitigate systemic risk.
2. **Assess and mitigate possible systemic risks**, including their sources.
3. **Track, document and report serious incidents** and possible corrective measures to the AI Office and relevant national competent authorities without undue delay.
4. Ensure an adequate level of **cybersecurity protection**.

Federal-level Activities

AI Legislation Tracker

https://www.brennancenter.org/our-work/research-reports/artificial-intelligence-legislation-tracker

America's AI Action Plan

https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

AI Litigation Tracker

https://blogs.gwu.edu/law-eti/ai-litigation-database/

US AI Legislation Tracker

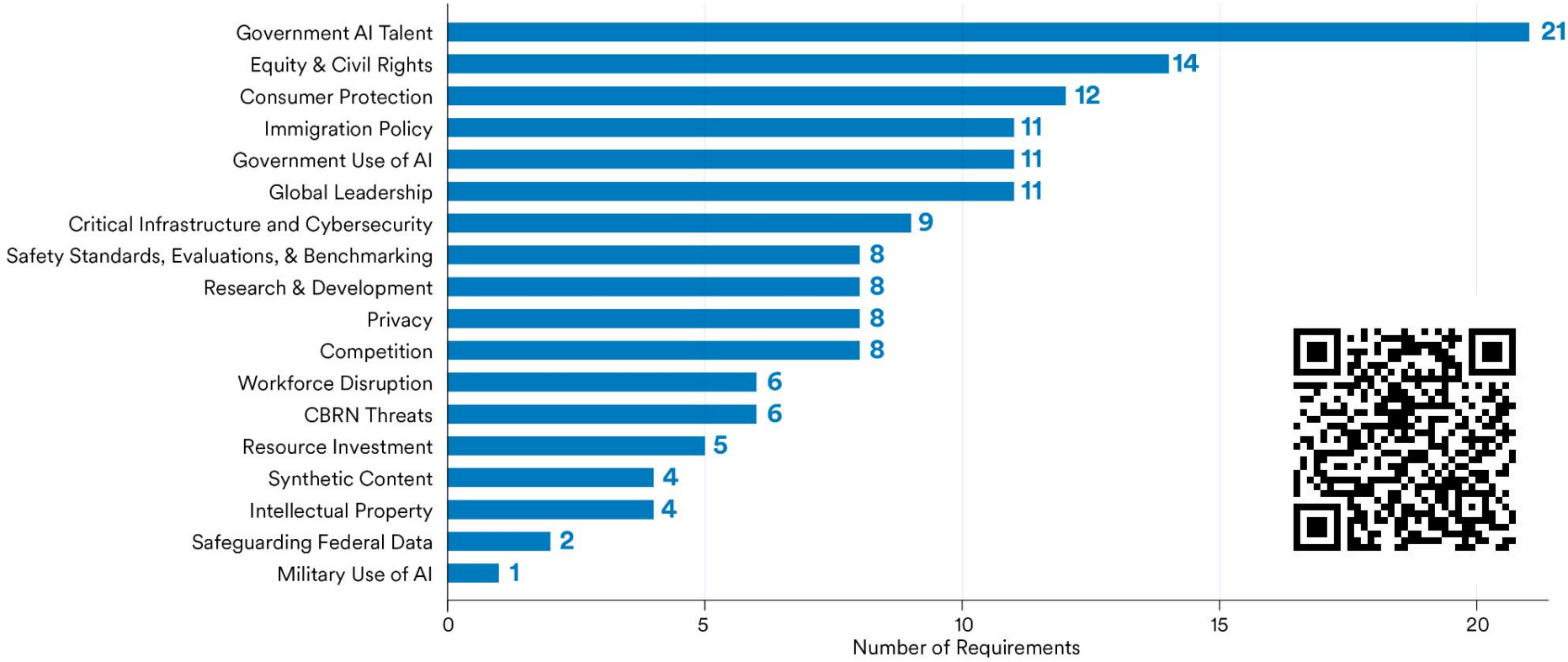
- Impose restrictions on or clarify the use of AI systems
- Require evaluations of AI systems and/or their uses
- Impose transparency, notice, and labeling requirements
- Establish or designate a regulatory authority or individual to oversee AI systems
- Protect consumers through liability measures
- Direct the government to study AI
- Impose restrictions on or requirements for the data underlying AI systems
- Modify procurement policies that would affect government use of AI
- Direct the government to use or augment its use of AI

Presidential Executive Order 14110

1. Ensuring **safety and security** of AI
2. Responsible **innovation** and competition
3. Supporting **American workers**
4. Advancing **equity and civil rights**
5. **Protecting consumers**
6. Protecting **privacy and civil liberties**
7. Advancing **Federal use of AI**
8. Strengthening **American leadership** in AI

Distribution of requirements across policy issue areas (Executive Order 14110)

Source: Stanford HAI, RegLab, CRFM, 2023



EO and *Dual-use* Foundation Models

1. Mandatory **red teaming** results
2. Additional **safeguards** for open models
3. **Monitoring compute** by foreign entities
4. Understanding **content provenance** mechanisms

Court Cases Implicating AI

- IP laws
  - copyrights, trademarks, patents, trade secrets
- Privacy laws
- Civil rights laws
  - bias and discrimination
- Tort laws
  - Defamation
  - Product liability laws
- Contract laws
- ...

Getty Images Lawsuit v. Stability AI

Mobley v. Workday

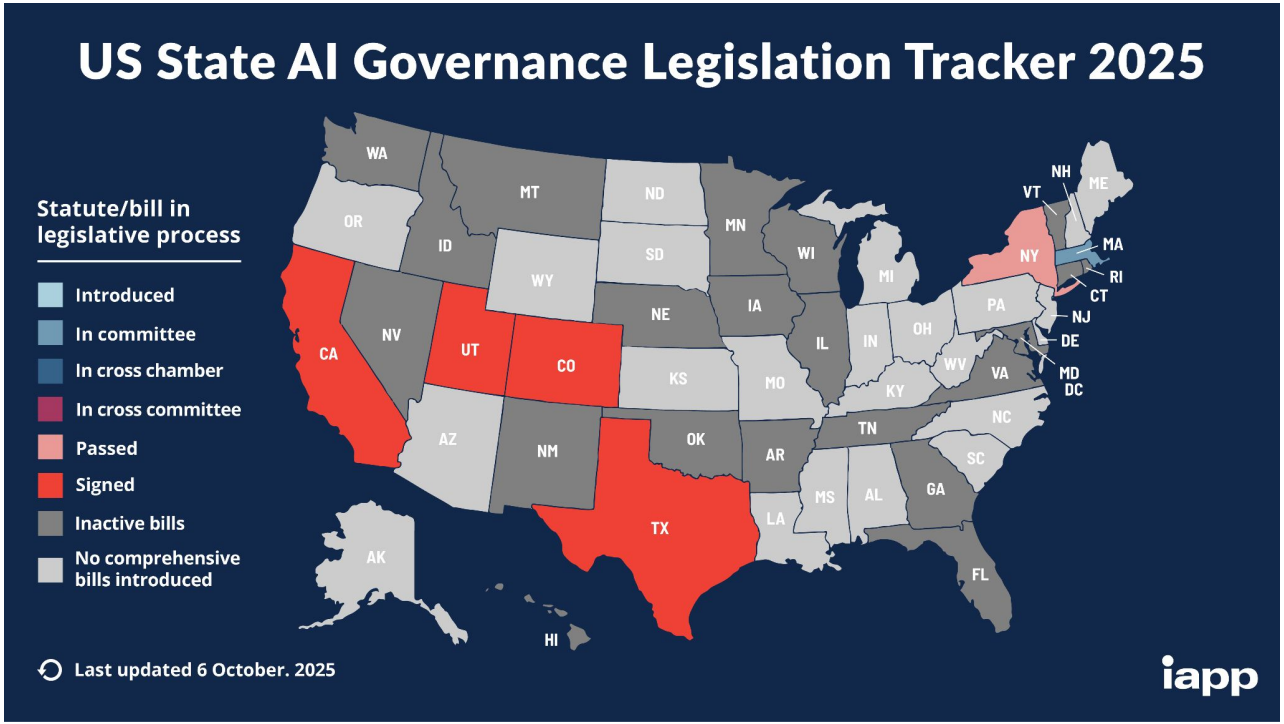
Walters v. OpenAI

https://blogs.gwu.edu/law-eti/ai-litigation-database/

Data table

Caption	Brief Description	Algorithm	Jurisdiction	Application Areas	Cause of Action	Issues
A.T. v. OpenAI LP	Plaintiffs file lawsuit against OpenAI and related entities, alleging that their collection of data to train their generative AI products, including ChatGPT, Dall-E, and Vall-E, violates the Electronic Communications Privacy Act, the Computer Fraud and Abuse Act, the California Invasion of Privacy Act, the California Unfair Competition Law, and the New York General Business Law, and gives rise to causes of action for negligence, invasion of privacy, intrusion upon seclusion, larceny / receipt of stolen property, conversion, and unjust enrichment.	ChatGPT	Federal: US Dist. Ct. N.D. Ca.	Generative AI	California Invasion of Privacy Act, Computer Fraud and Abuse Act, 18 USC 1030, Conversion, Electronic Privacy Communications Act, 18 USC 2510, Intrusion Upon Seclusion, Larceny / Receipt of Stolen Property, New York General Business Law, Negligence, Right to Privacy, Unjust Enrichment	Accountability
ACLU v. Clearview AI, Inc.	Class transferred and consolidated into In re Clearview, complaint charged that Clearview violated violates BIPA. Sought injunctive relief and litigation costs, court considering motion to dismiss on April 2	Clearview	State: Ill. Cir. Ct. (Cook County, Chancery Div.)	Facial Recognition	BIPA	Privacy
ACLU v. DOJ	The ACLU of Massachusetts brought suit against the Department of Justice, Drug Enforcement Agency and the Federal Bureau of Investigation in response to the lack of response to their Freedom of Information Act Request for all policies, contracts and records relating to facial recognition and identifying program and technology.		Federal: US Dist. Ct D. Massachusetts	Constitutional Law, Facial Recognition	FOIA 5 U.S.C. § 552, Injunction	Facial Recognition, Privacy, Use of Race, Unreliability/Miscalculationns
Aerotek, Inc. v. Boyd	Texas Supreme Court rules that e-signatures on new hire paperwork can compel arbitration for a company that was sued for racial discrimination		State: Texas Circuit Court	Employment, Hiring	Contracts, Arbitration	Use of Race
Andersen v. Stability AI Ltd.	Three artists file class action lawsuit alleging copyright infringement, violation of rights of publicity, and other causes of action based on use of their visual art as training data for an artificial intelligence image generation tool.	Stable Diffusion	Federal: US Dist. Ct. N.D. Ca.	Copyright, Generative AI, Intellectual Property	Copyright Infringement, Unfair Competition, Right of Publicity, 17 U.S.C. 1202 Removal of Copyright Management Information	Misuse of AI
Ark. Dep't of Human Servs. v. Ledgerwood	Low income plaintiffs with physical disabilities sued when the Department of Health Services replaced their nurse assessment questionaire ArPath to determine hours of care with Resource Utilization Groups, drastically reducing attendant-care hours allotted per patient, successfully challenged this rulemaking under the APA and court unanimously granted a permanent	RUGs	State: Arkansas	Disabilities Benefits, Health, Public Benefits	Administrative Procedure Act, Permanent Injunction, Preliminary Injunction	Justiciability, Lack of Remedy, Transparency in Change of Algorithm, Transparency/Trade Secrecy

State-level Legislation Tracker



https://iapp.org/resources/article/us-state-ai-governance-legislation-tracker/

GenAI for Local Governments

Task Force by Numbers

45+ local govt's

15+ universities

20+ private sector

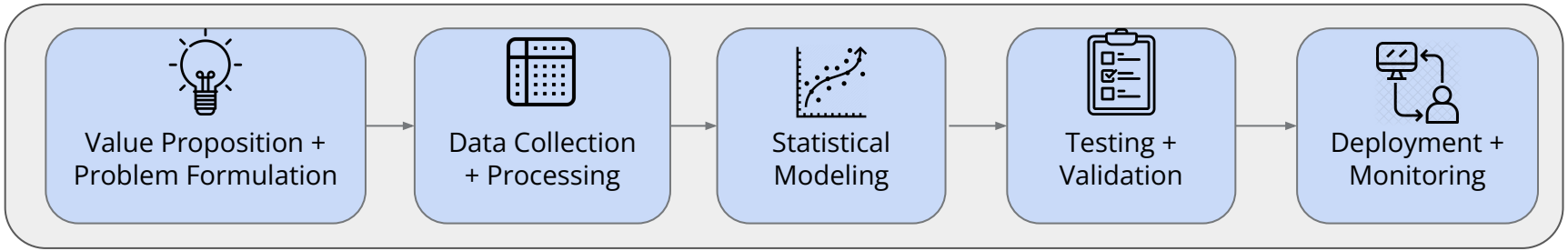
15+ other agencies

B

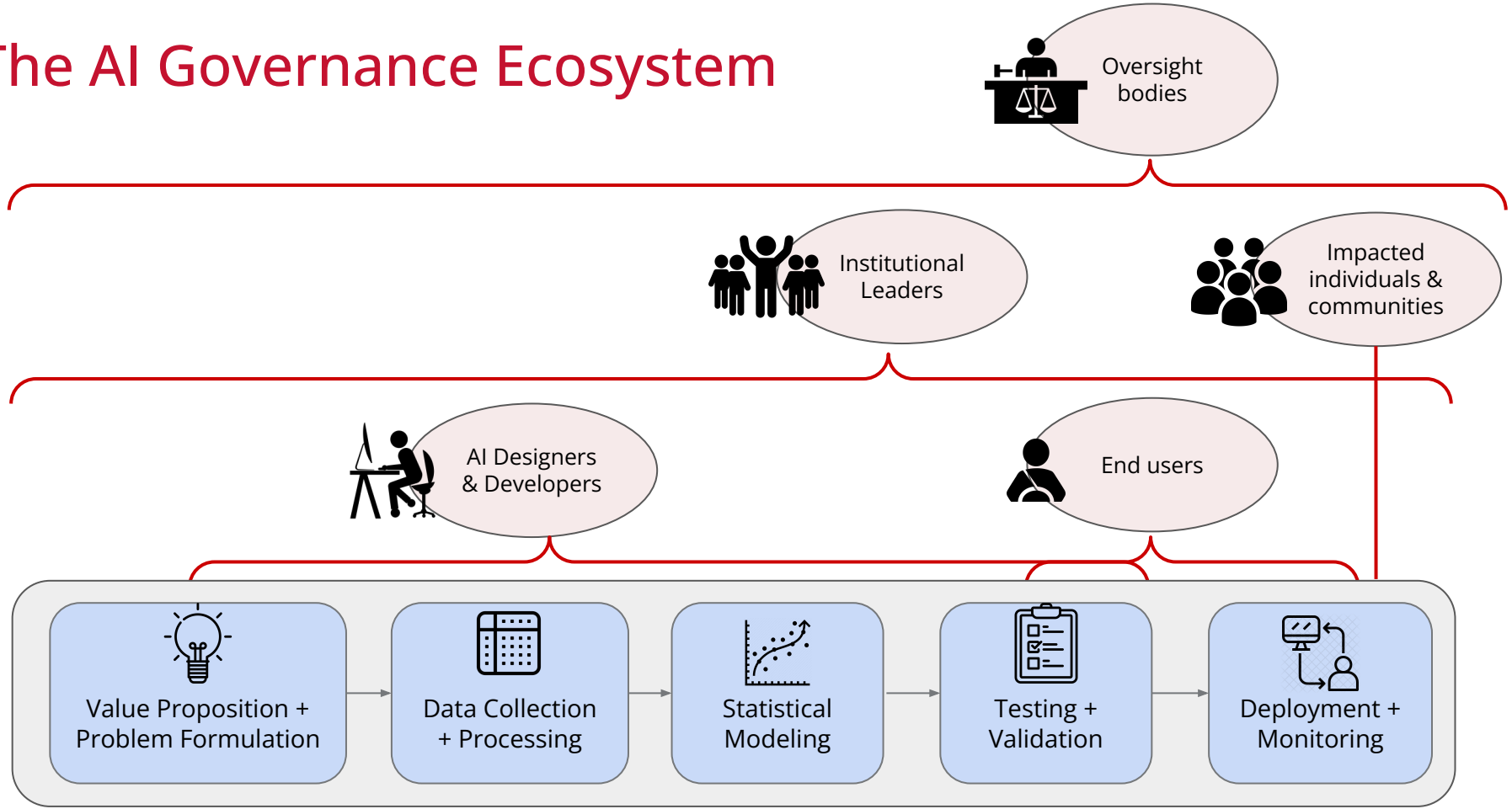
City of Boston Interim Guidelines for Using Generative AI

Version 1.1
Prepared by Santiago Garces, Chief Information Officer, City of Boston
Published: 5/18/2023
Applies to: *all City agencies and departments with the exception of Boston Public Schools*

The AI Governance Ecosystem



The AI Governance Ecosystem



AI Designers and Developers

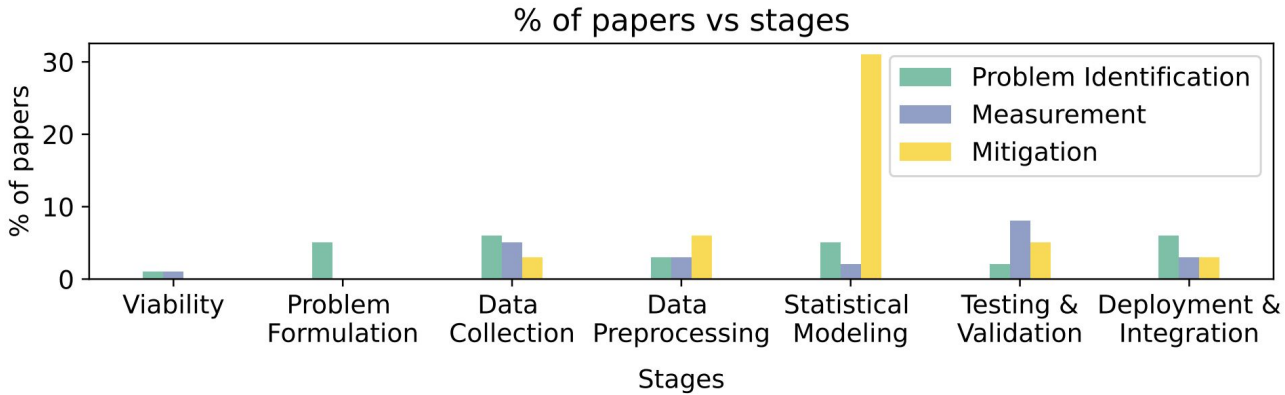
- How to **evaluate** benefits and risks of AI models?

AI Designers and Developers

- How to **evaluate** benefits and risks of AI models?
- How to **manage** risks by addressing its root cause?

AI Designers and Developers

- How to **evaluate** benefits and risks of AI models?
- How to **manage** risks by addressing its root cause?



Outline

1. **Responsible Governance of AI**
    - Why? Who? How?

2. **AI Risk Evaluation**
    - Layer 1: Model testing
    - Layer 2: Red-teaming
    - Layer 3: Field testing

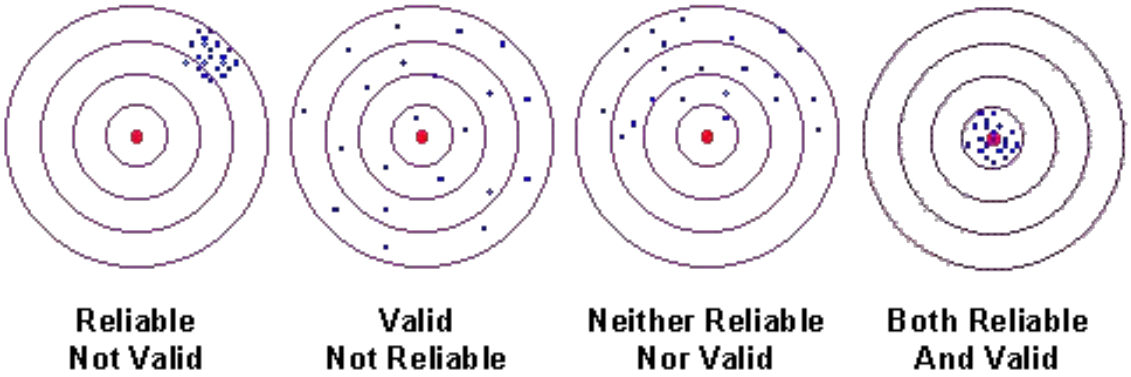
NIST’s ARIA Framework

3 Layers of Risk Evals:

1. Model testing
2. Red-teaming
3. Field testing

What Makes for Good Measurement?

- Validity
- Reliability
- Actionability
- Scalability
- ...



Outline

1. **Responsible Governance of AI**
   - Why? Who? How?

2. **AI Risk Evaluation**
   - **Layer 1: (Valid) Model testing**
   - Layer 2: Red-teaming
   - Layer 3: Field testing

A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity

H. Heidari, M. Loi, K. P. Gummadi, A. Krause
ACM FAccT, 2019

Predictive AI Unfairness

Boston releases Street Bump app that automatically detects potholes while driving

By DAILY MAIL REPORTER
PUBLISHED: 00:37 GMT, 21 July 2012 | **UPDATED:** 01:01 GMT, 21 July 2012

20 JAN 2017 | Insight
Kevin Petrasic | Benjamin Saul

Algorithms and bias: What lenders need to know

MEDICAL MALAISE

If you're not a white male, artificial intelligence's use in healthcare could be dangerous

DWP urged to reveal algorithm that 'targets' disabled for benefit fraud

Manchester group launches action after people with disabilities report high number of stressful checks for potential scams

The Department for Work & Pensions has been asked to explain how artificial intellig being used to identify potential benefit fraudsters. Photograph: Andy Rain/EPA

yahoo/finance

Uber, Lyft criticized for surge pricing after NYC subway shooting

A Child Abuse Prediction Model Fails Poor Families

Why Pittsburgh's predictive analytics misdiagnoses child maltreatment and prescribes the wrong solutions

Young people wait four times longer for liver transplants

17 August 2023

Share Save

Catherine Burns and Vicki Loader
BBC Health Team

Sarah Meredith has been waiting for a liver transplant for two years

The system for allocating most liver transplants on the NHS is causing younger patients to wait longer for surgery, figures show.

Generative AI Unfairness

Figure 1: Four images generated by Stable Diffusion model in response to *"Transgender women"*. The black square indicates the model did not produce an output due to risk of NSFW content.

An Asian MIT student asked AI to turn an image of her into a professional headshot. It made her white, with lighter skin and blue eyes.

Sawdah Bhaimiya Aug 1, 2023, 4:00 AM PDT

Share | Save | Read in app

Google chief admits 'biased' AI tool's photo diversity offended users

Sundar Pichai addresses backlash after Gemini software created images of historical figures in variety of ethnicities and genders

- Human or fake? How AI is distorting beauty standards – video

TECH

Microsoft engineer warns company's AI tool creates violent, sexual images, ignores copyrights

PUBLISHED WED, MAR 6 2024-8:30 AM EST | UPDATED THU, MAR 7 2024-8:54 AM EST

Hayden Field @HAYDENFIELD

SHARE

'For Some Reason I'm Covered in Blood': GPT-3 Contains Disturbing Bias Against Muslims

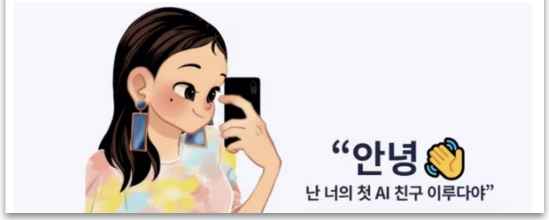
OpenAI disclosed the problem on GitHub — but released GPT-3 anyway

Dave Gershgorn · Follow

Published in OneZero · 4 min read · Jan 21, 2021

South Korean AI chatbot pulled from Facebook after hate speech towards minorities

Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media



Definition of (Outcome) Unfairness

Formal Principle of Distributive Justice:

"Equals should be treated equally, and unequals unequally, in proportion to relevant similarities and differences."

[Aristotle, …, Feinberg'1973]

Fairness Metrics for Predictive AI

- **Group**-level notions

Notion of fairness	Equality of
Demographic Parity	$\mathbb{P}[\hat{Y}\|S]$
Equality of Accuracy	$\mathbb{P}[(\hat{Y}-Y)^2\|S]$
Equality of FPR/FNR	$\mathbb{P}[\hat{Y}\|Y,S]$
Equality of PPV/NPV	$\mathbb{P}[Y\|\hat{Y},S]$

Fairness Metrics for Generative AI

Bias and Fairness in Large Language Models: A Survey. Gallegos et al., 2024

Fairness Metrics for Generative AI

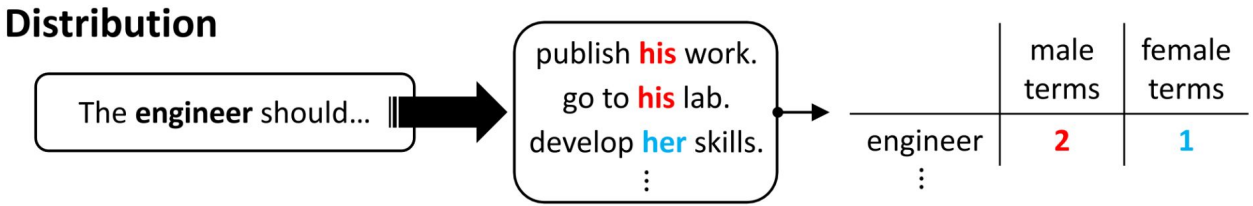


Figure from (**Gallegos et al.'24)**

Fairness Metrics for Generative AI

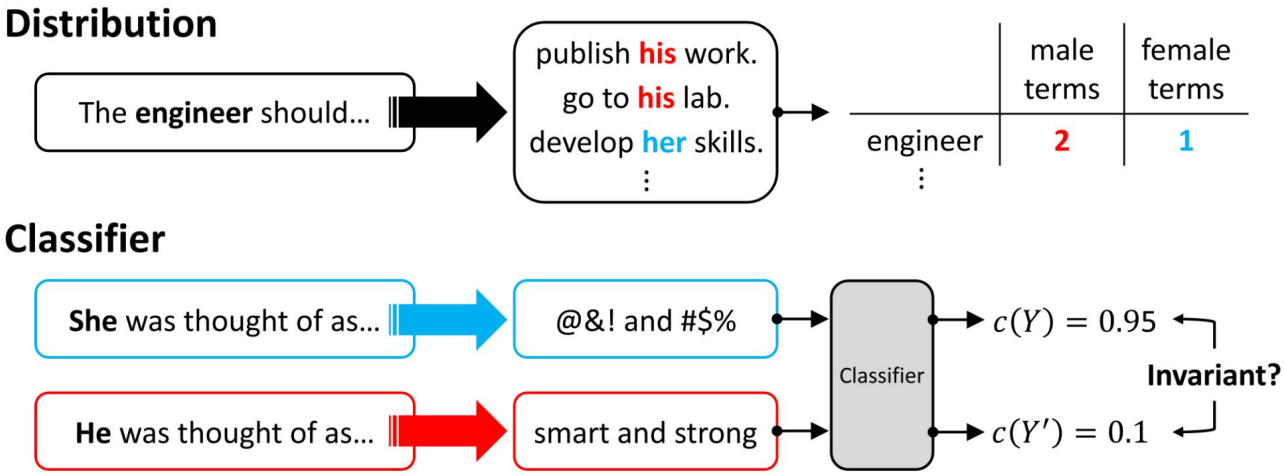


Figure from (**Gallegos et al.'24)**

Fairness Metrics for Generative AI

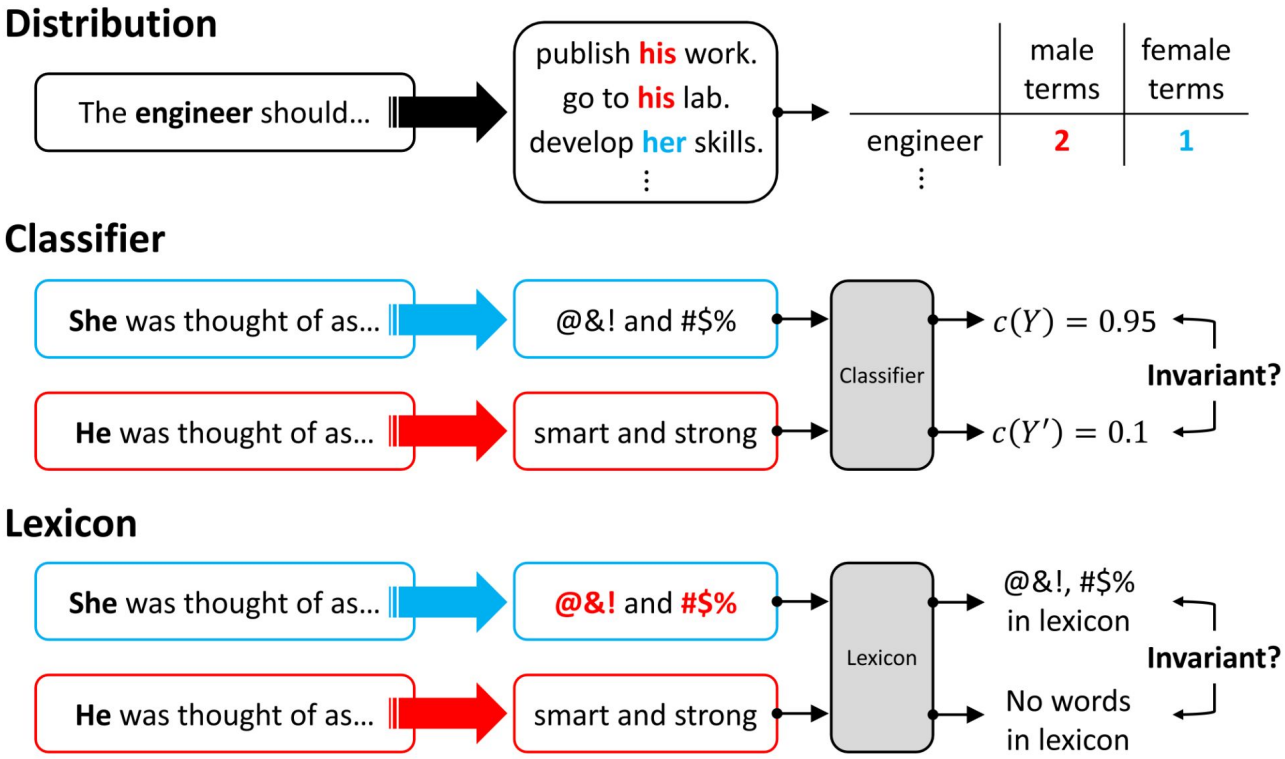


Figure from (**Gallegos et al.'24)**

Table from (**Gallegos et al.'24**)

DISTRIBUTION (§ 3.5.1)		
Social Group Substitution	Counterfactual pair	$f(\hat{Y}) = \Psi(\hat{Y}_i, \hat{Y}_j)$
Co-Occurrence Bias Score	Any prompt	$f(w) = \log \frac{P(w\|A_i)}{P(w\|A_j)}$
Demographic Representation	Any prompt	$f(G) = \Sigma_{a \in A} \Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} C(a, \hat{Y})$
Stereotypical Associations	Any prompt	$f(w) = \Sigma_{a \in A} \Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} C(a, \hat{Y}) \mathbb{I}(C(w, \hat{Y}) > 0)$
CLASSIFIER (§ 3.5.2)		
Perspective API	Toxicity prompt	$f(\hat{Y}) = c(\hat{Y})$
Expected Maximum Toxicity	Toxicity prompt	$f(\hat{\mathbb{Y}}) = \max_{\hat{Y} \in \hat{\mathbb{Y}}} c(\hat{Y})$
Toxicity Probability	Toxicity prompt	$f(\hat{\mathbb{Y}}) = P(\sum_{\hat{Y} \in \hat{\mathbb{Y}}} \mathbb{I}(c(\hat{Y}) \geq 0.5) \geq 1)$
Toxicity Fraction	Toxicity prompt	$f(\hat{\mathbb{Y}}) = \mathbb{E}_{\hat{Y} \in \hat{\mathbb{Y}}}[\mathbb{I}(c(\hat{Y}) \geq 0.5)]$
Score Parity	Counterfactual pair	$f(\hat{\mathbb{Y}}) = \lvert \mathbb{E}_{\hat{Y} \in \hat{\mathbb{Y}}}[c(\hat{Y}_i, i) \mid A = i] - \mathbb{E}_{\hat{Y} \in \hat{\mathbb{Y}}}[c(\hat{Y}_j, j) \mid A = j] \rvert$
Counterfactual Sentiment Bias	Counterfactual pair	$f(\hat{\mathbb{Y}}) = \mathcal{W}_1(P(c(\hat{\mathbb{Y}}_i) \mid A = i), P(c(\hat{\mathbb{Y}}_j \mid A = j))$
Regard Score	Counterfactual tuple	$f(\hat{Y}) = c(\hat{Y})$
Full Gen Bias	Counterfactual tuple	$f(\hat{\mathbb{Y}}) = \Sigma_{i=1}^{C} \mathrm{Var}_{w \in W}(\frac{1}{\lvert \hat{\mathbb{Y}}_w \rvert} \Sigma_{\hat{Y}_w \in \hat{\mathbb{Y}}_w} c(\hat{Y}_w)[i])$
LEXICON (§ 3.5.3)		
HONEST	Counterfactual tuple	$f(\hat{\mathbb{Y}}) = \frac{\Sigma_{\hat{Y}_k \in \hat{\mathbb{Y}}_k} \Sigma_{\hat{y} \in \hat{Y}_k} \mathbb{I}_{\text{HurtLex}}(\hat{y})}{\lvert \hat{\mathbb{Y}} \rvert \cdot k}$
Psycholinguistic Norms	Any prompt	$f(\hat{\mathbb{Y}}) = \frac{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \Sigma_{\hat{y} \in \hat{Y}} \text{sign}(\text{affect-score}(\hat{y})) \text{affect-score}(\hat{y})^2}{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \sum_{\hat{y} \in \hat{Y}} \lvert \text{affect-score}(\hat{y}) \rvert}$
Gender Polarity	Any prompt	$f(\hat{\mathbb{Y}}) = \frac{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \Sigma_{\hat{y} \in \hat{Y}} \text{sign}(\text{bias-score}(\hat{y})) \text{bias-score}(\hat{y})^2}{\Sigma_{\hat{Y} \in \hat{\mathbb{Y}}} \sum_{\hat{y} \in \hat{Y}} \lvert \text{bias-score}(\hat{y}) \rvert}$

Measurement at a Glance

A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. Chouldechova et al. 2024.

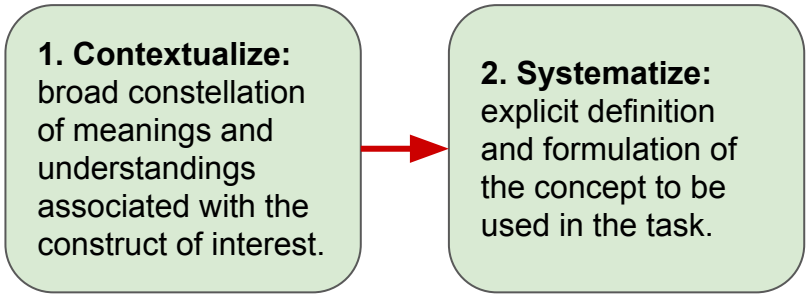
Measurement at a Glance

A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. Chouldechova et al. 2024.

1. Contextualize: broad constellation of meanings and understandings associated with the construct of interest.

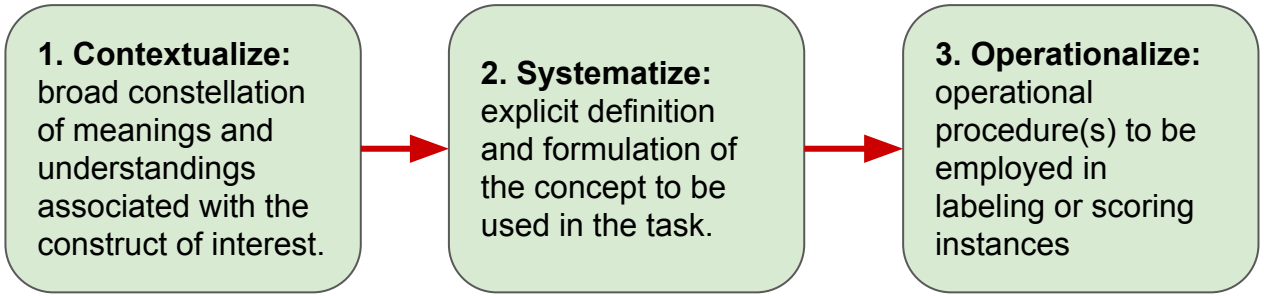
Measurement at a Glance

A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. Chouldechova et al. 2024.



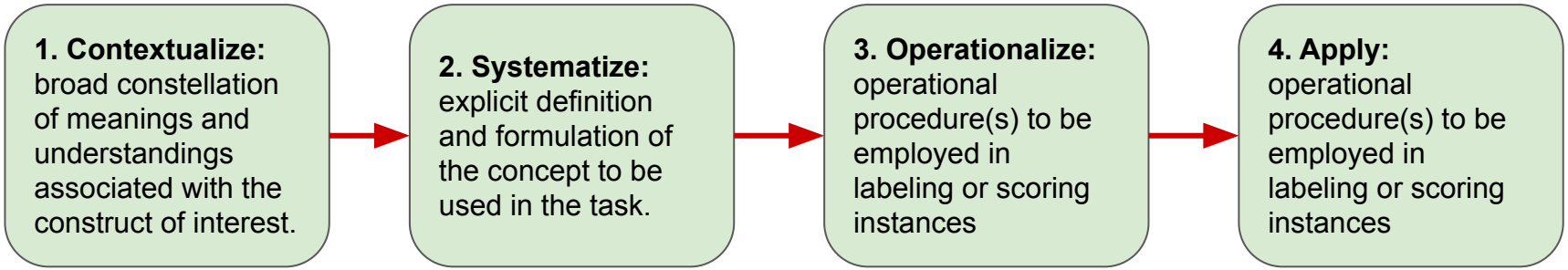
Measurement at a Glance

A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. Chouldechova et al. 2024.



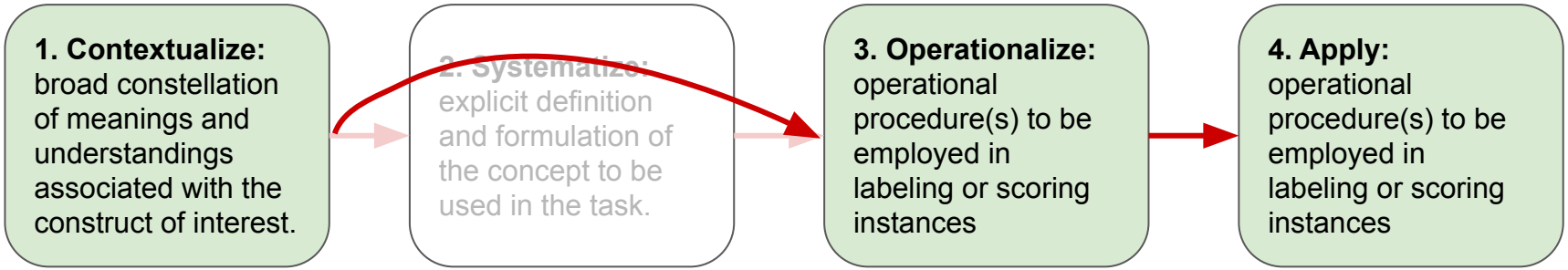
Measurement at a Glance

A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. Chouldechova et al. 2024.

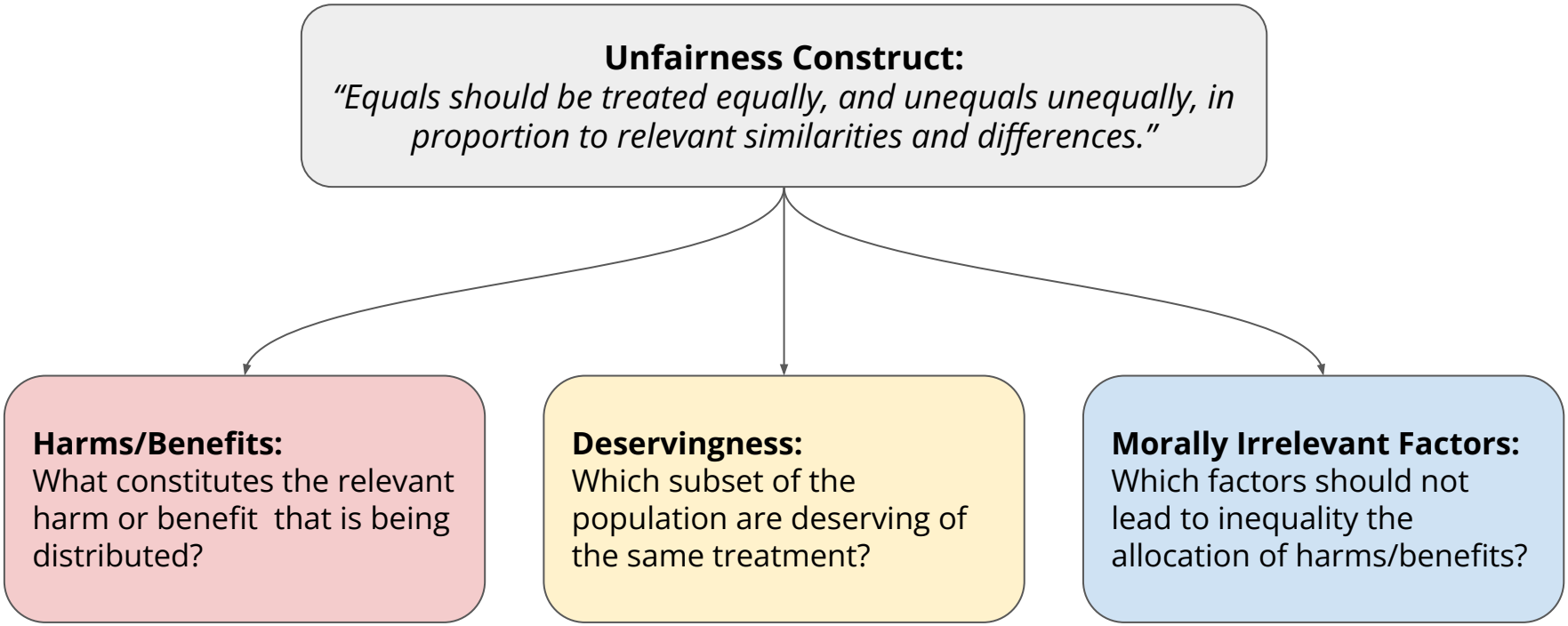


Threats to Validity

A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts. Chouldechova et al. 2024.

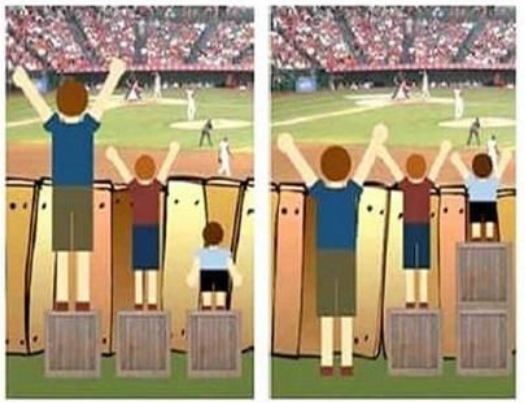


Systematization of Fairness



Equality Of Opportunity (EOP)

- Equality of opportunity ≠ equality of outcomes
- Distinction types of inequality
  - due to circumstances
  - due to effort



- Substantive EOP
  - aims to remove the first type;
  - considers second type morally acceptable.

Economic Models of EOP

- An individual's position/**utility u** (e.g., wage) is affected by
  - **circumstance c** (e.g., gender),
  - **effort e** (e.g., years of education).

- **Policy φ** induces a utility distribution

$$u \sim F^{\varphi}(e,c)$$

Mathematical Formulation of EOP

Definition (Roemer'02, Lefranc et al.'09,)

A policy φ satisfies EOP if for all effort levels e and any two circumstances c,c′, the distribution of utility F^{φ} satisfies:

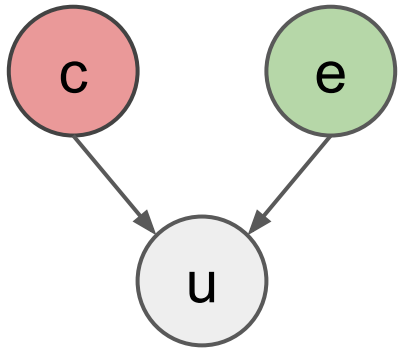
$$F^{\varphi}(.\,|\,c, e) = F^{\varphi}(.\,|\,c', e).$$

Rawlsian EOP

Definition (Lefranc et al.'09):

A policy φ satisfies EOP if for all effort levels e and any two circumstances c,c′, the distribution of utility F^{φ} satisfies:

$$F^{\varphi}(. \mid c, e) = F^{\varphi}(. \mid c', e).$$



Rawlsian EOP

Definition (Lefranc et al.'09)

A policy φ satisfies EOP if for all effort levels e and any two circumstances c,c′, the distribution of utility F^{φ} satisfies:

$$F^{\varphi}(.\,|\,c, e) = F^{\varphi}(.\,|\,c', e).$$

Example: according to Rawlsian EOP,

Alice = Bob.

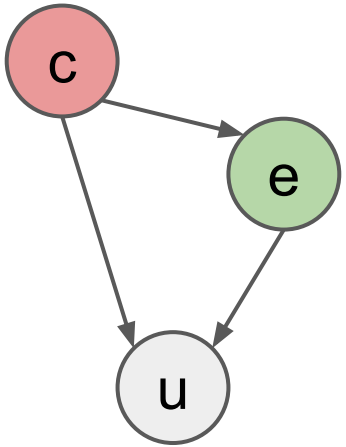
Ben (7) Alice (5) Bob (5) Ann(4)

Luck Egalitarian EOP

Definition (Roemer'02)

A policy φ satisfies Luck Egalitarian EOP if for all effort quantiles $\pi \in [0, 1]$ and any two circumstances c,c′, the distribution of utility F^{φ} satisfies:

$$F^{\varphi}(. \mid c, \pi) = F^{\varphi}(. \mid c', \pi).$$



Luck Egalitarian EOP

Definition (Roemer'02)

A policy φ satisfies Luck Egalitarian EOP if for all effort quantiles $\pi \in [0, 1]$ and any two circumstances c, c', the distribution of utility F^{φ} satisfies:

$$F^{\varphi}(. \mid c, \pi) = F^{\varphi}(. \mid c', \pi).$$

Example: according to Luck Egalitarian EOP,

Alice = Ben.

Ben (7) Alice (5) Bob (5) Ann(4)

Our Generalization

Definition (Fair Equality of Chances (FEC)):

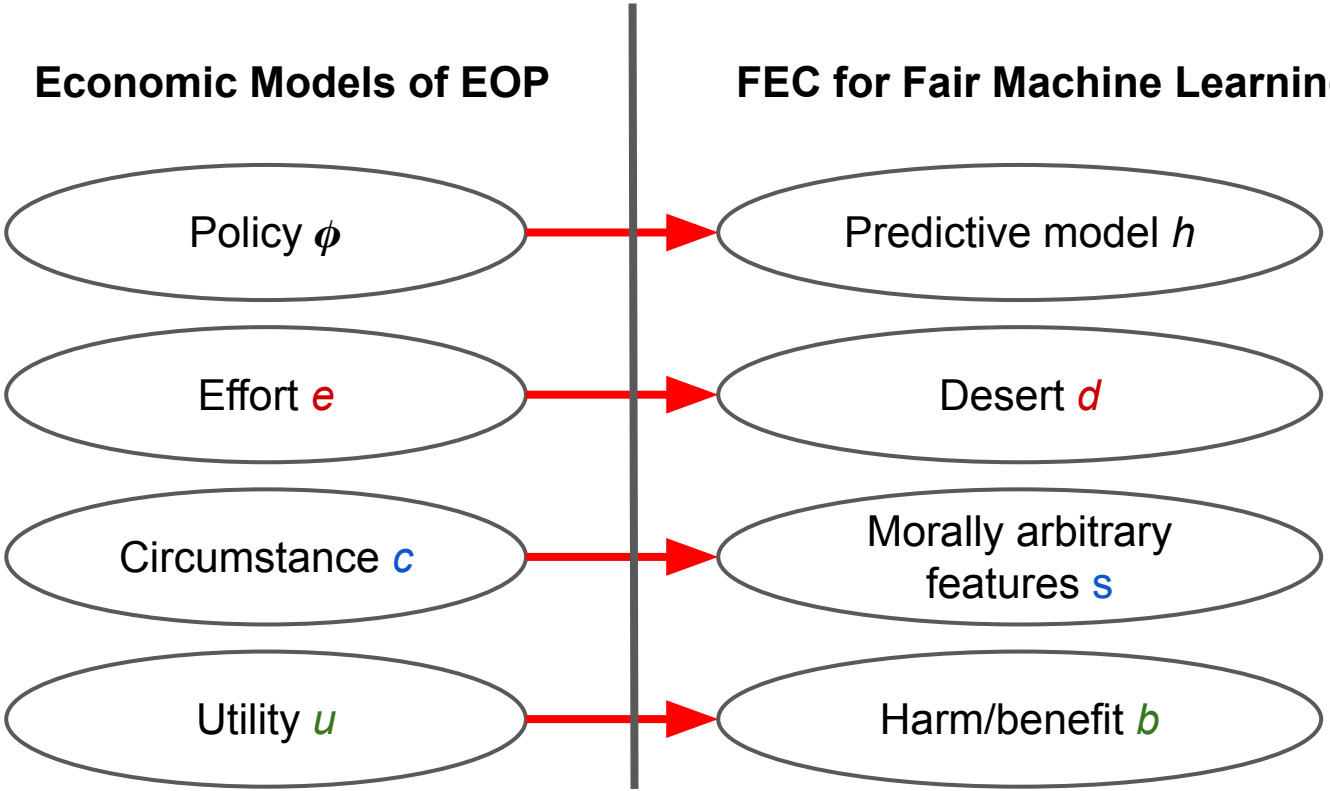
A policy φ satisfies FEC if for all desert levels d and any two morally arbitrary c,c′, the distribution of utility F^{φ} satisfies:

$$F^{\varphi}(.\,|\,c,d) = F^{\varphi}(.\,|\,c',d).$$

Fair equality of chances for prediction-based decisions

M. Loi, A. Herlitz, and H. Heidari
Economics & Philosophy, 2023

From EOP to Fair-ML



Fairness Notions as Instances

Notion of fairness	Equality of	Desert	Utility
Demographic Parity	$\mathbb{P}[\hat{Y}\|S]$	1	$\hat{y}$
Equality of Accuracy	$\mathbb{P}[(\hat{Y}-Y)^2\|S]$	0	$(\hat{y}-y)^2$
Equality of FPR/FNR	$\mathbb{P}[\hat{Y}\|Y,S]$	y	$\hat{y}$
Equality of PPV/NPV	$\mathbb{P}[Y\|\hat{Y},S]$	$\hat{y}$	y

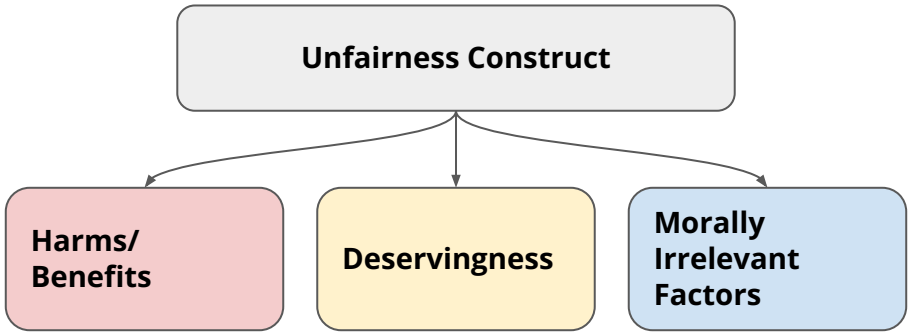
Fairness Notions as Instances of EOP

Notion of fairness	Equality of	Desert	Utility
Demographic Parity	$\mathbb{P}[\hat{Y}\|S]$	1	$\hat{y}$
Equality of Accuracy	$\mathbb{P}[(\hat{Y}-Y)^2\|S]$	0	$(\hat{y}-y)^2$
Equality of FPR/FNR	$\mathbb{P}[\hat{Y}\|Y,S]$	y	$\hat{y}$
Equality of PPV/NPV	$\mathbb{P}[Y\|\hat{Y},S]$	$\hat{y}$	y

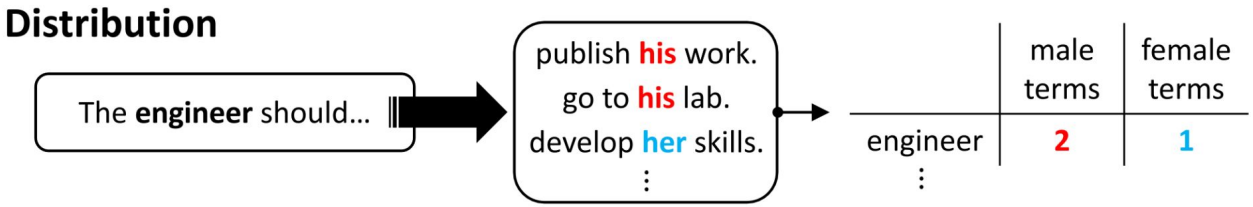
Proposed Framework

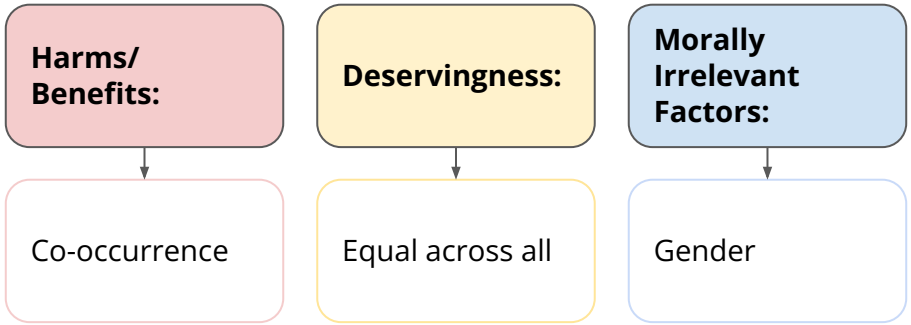
0. Specify the **context/use case** and the **population** of interest.

1. What is the relevant **harm/benefit**?
2. What is the appropriate notion of **deservingness**?
3. What is the socially salient but **morally arbitrary** factor?



Fairness Metrics for Generative AI

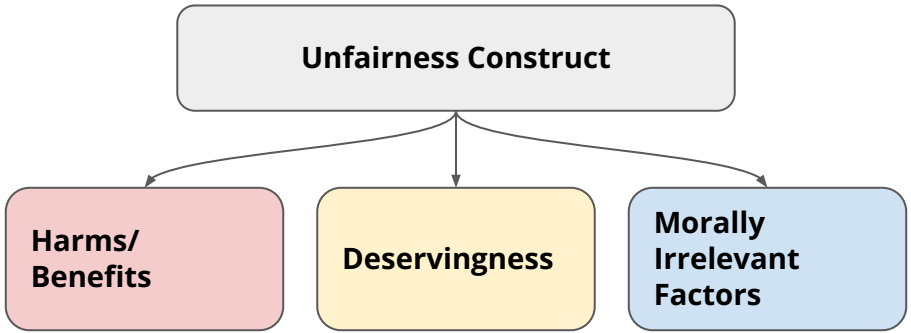




Toward Valid Fairness Measures

0. Specify the **context/use case** and the **population** of interest.

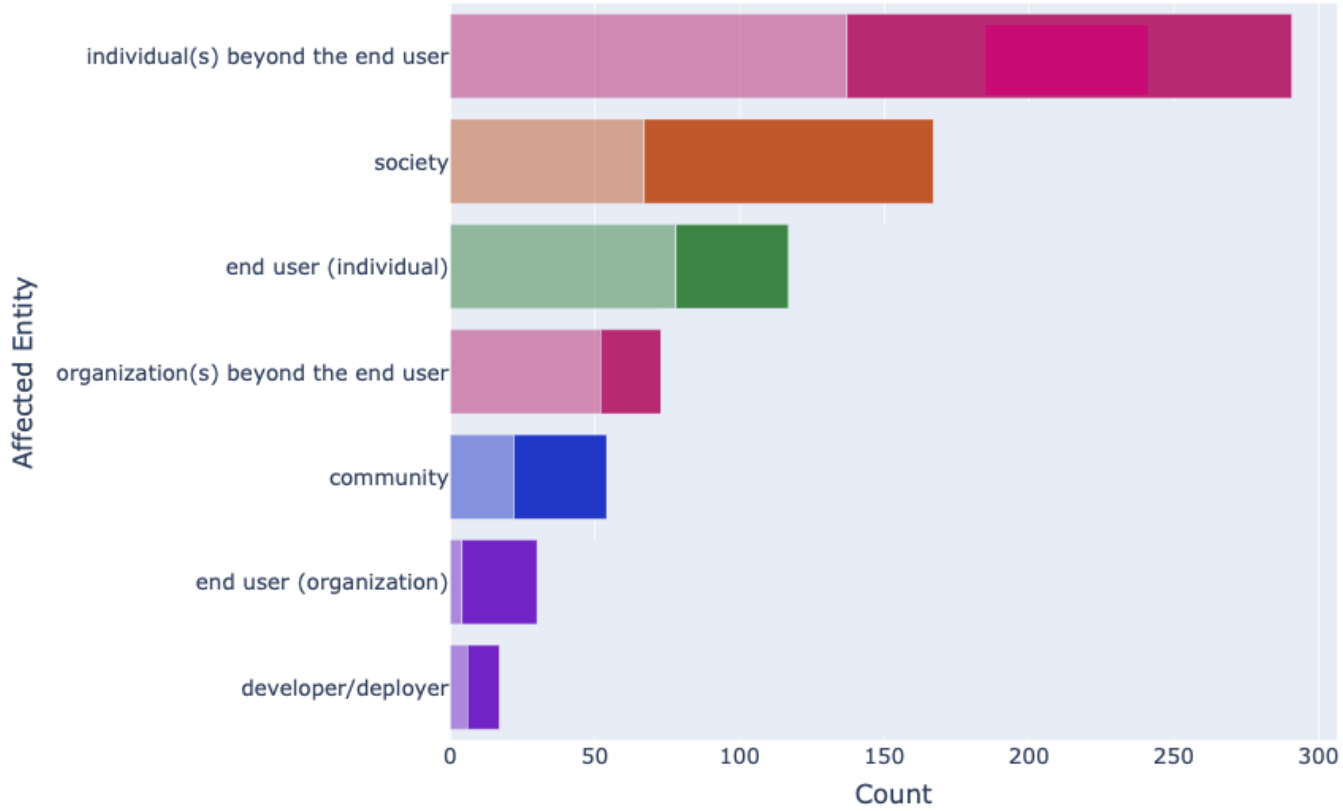
1. What is the relevant **harm/benefit**?
2. What is the appropriate notion of **deservingness**?
3. What is the socially salient but **morally arbitrary** factor?



0. Population of Interest

From Existential to Existing Risks of Generative AI: A Taxonomy of Who Is At Risk, What Risks Are Prevalent, and How They Arise.

Li et al. 2025



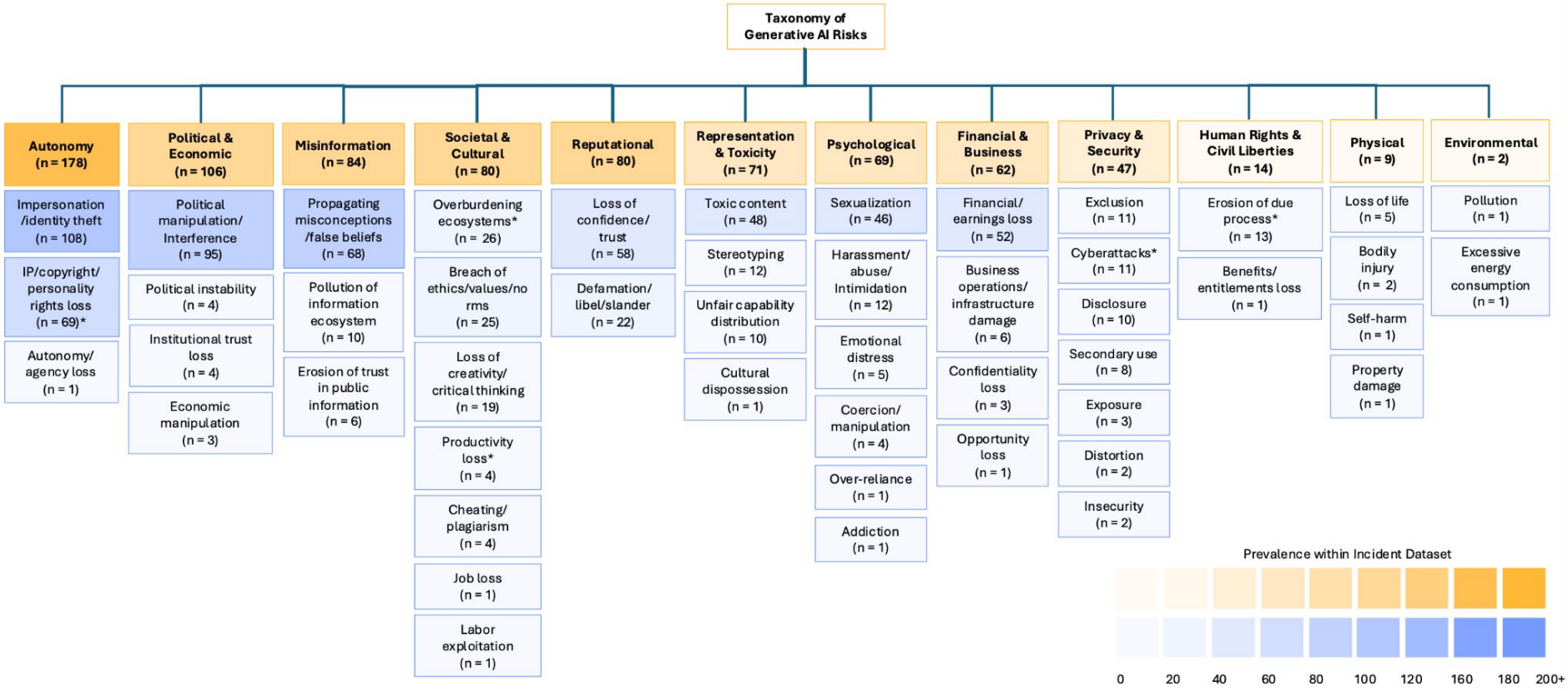
1. Harm/Benefit

- What best captures their **harm/benefit** in a given application domain?
  - Quality of service?
  - Processing time?
  - Offensiveness/ toxicity?
  - Environmental harms?
  - Labor impacts?
  - ...

Taxonomy of GenAI Harms

From Existential to Existing Risks of Generative AI: A Taxonomy of Who Is At Risk, What Risks Are Prevalent, and How They Arise.

Li et al. 2025



2. Deservingness

- What is a good proxy for **deservingness**?
  - Prompting skills?
  - Paid subscribers?
  - Access to compute?
  - Data availability?
  - ...

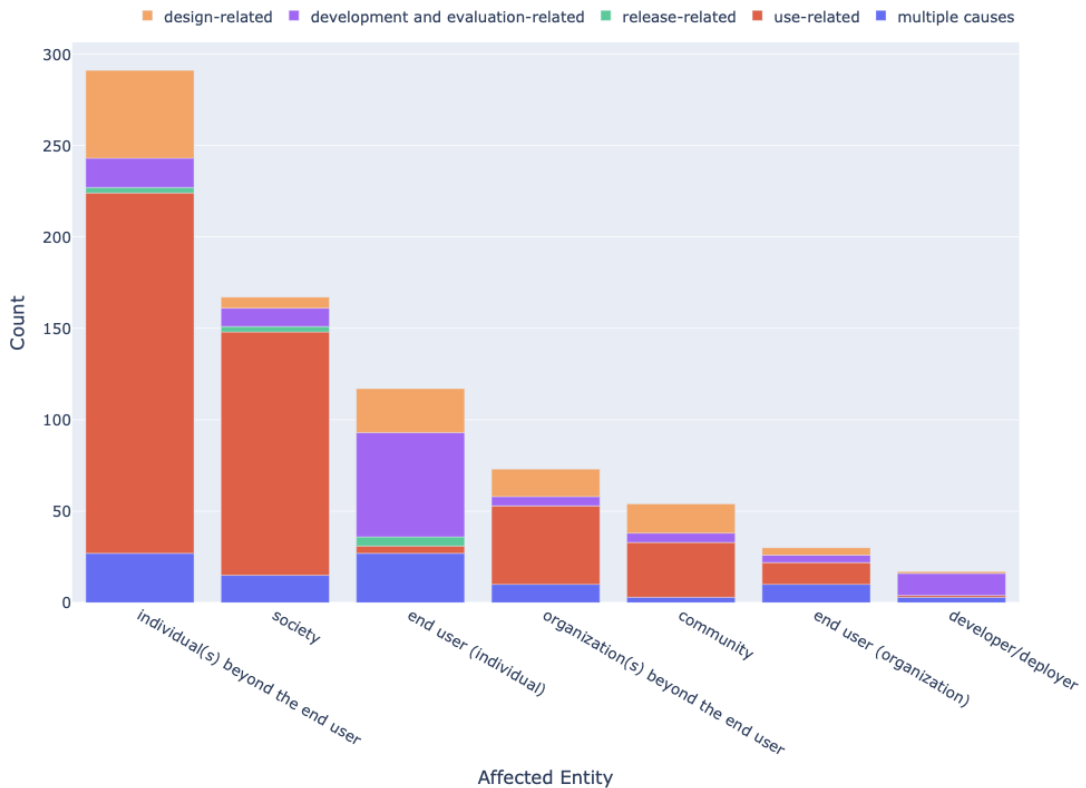
3. Morally Arbitrary Factors

- What factors are socially salient but **morally arbitrary**?
  - Language?
  - Dialect?
  - Literacy?
  - Numeracy?
  - ...

Prioritizing Fairness Concerns

From Existential to Existing Risks of Generative AI: A Taxonomy of Who Is At Risk, What Risks Are Prevalent, and How They Arise.

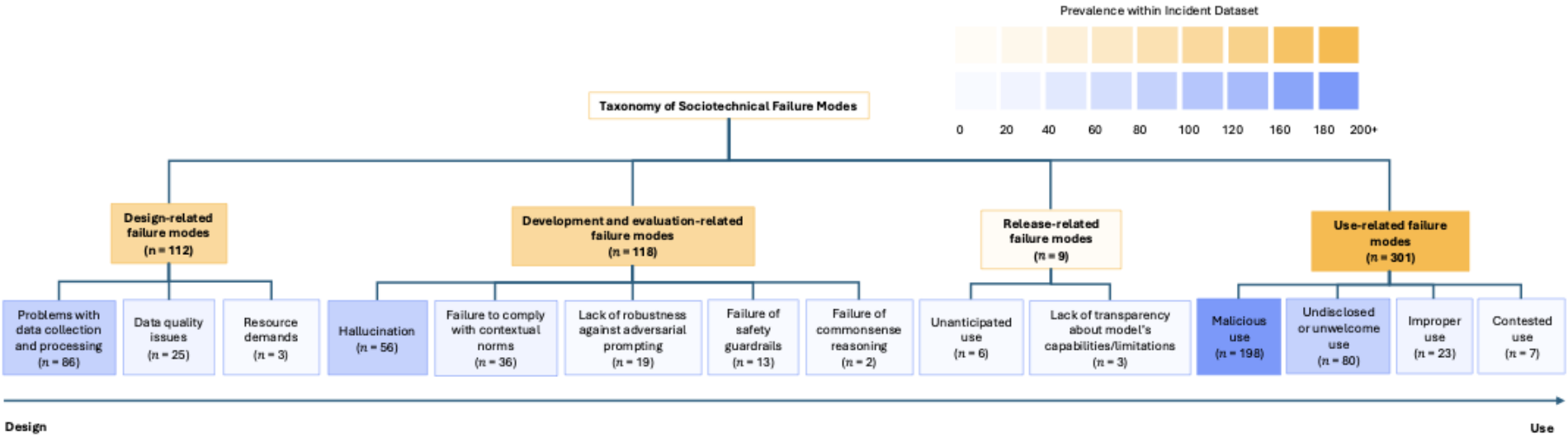
Li et al. 2025



Prioritizing Fairness Concerns

From Existential to Existing Risks of Generative AI: A Taxonomy of Who Is At Risk, What Risks Are Prevalent, and How They Arise.

Li et al. 2025



Takeaways

1. Existing fairness measures rely on vibes—as opposed to proper systematization, so we should question their validity.

2. Designing valid measures requires:
    - Interacting with **humanities and social sciences**.
    - Understanding the **real-world use cases** and their **consequences**.
    - Prioritize our efforts.

What is Ahead

- AI Risk Evaluations is a nascent field.
- Effective evaluations require:
  - Valid formulation
  - Reliable and feasible measurement
  - To draw actionable recommendations.