

Winter Term 21/22

# Artificial Intelligence, Ethics & Engineering

## Lecture-2: Towards Responsible AI - Principles and Criteria for Fairness

Prof. Dr. Holger Giese ([holger.giese@hpi.uni-potsdam.de](mailto:holger.giese@hpi.uni-potsdam.de))

Christian Medeiros Adriano ([christian.adriano@hpi.de](mailto:christian.adriano@hpi.de)) - **"Chris"**

Christian Zöllner ([Christian.zoellner@hpi.de](mailto:Christian.zoellner@hpi.de))

# Project Example: Engineering support of Design, Verification & Validation of Ethical Argumentation

---



**Outcome:** Methodology, Models, Data, and Tool Prototype

**Topic:** Perception of Competing Argumentations for Ethical Dilemmas on <Fairness x Trustworthiness, Privacy x Safety, etc.>

**Domains/Ethical Dilemma:** Recommender Systems, Social Networks, Surveillance and Identification Systems, Medical diagnosis

## **Possible Project Tasks:**

- Describe the specific ethical dilemma with examples (domain-specific or multiple-domains)
- Describe the arguments that cover the various aspects of the dilemma (use multiple definitions/understandings of an ethical principle)
- Codify arguments using some model from a methodology or tool
- Use the model to generate a critique on the dilemma, e.g., identify fallacies, false assumptions, misunderstandings, tc.
- Design an experiment to evaluate the human subjective perception of these arguments (use threats to validity to check your assumptions)
- Revise the design after executing a pilot of the experiment (colleagues)
- Run the large-scale experiment

# Project Example: Engineering support of Design, Verification & Validation of Ethical Argumentation

---



## Discussion of Findings:

- Which fallacies or false assumptions were successfully identified? Which were not?
- Which counterarguments for these fallacies are well perceived? By whom?
- Are there external factors (demographics) correlated with the findings?
- Is there any detectable confounding? Or plausibly hidden?

## Implications to Engineering:

- What variations in the argumentations could be credited/blamed for a more positive/negative perception?
- Are there possible improvements in the model that codifies the arguments? Fallacies that were not identified?
- Are there plausible ways to intervene in the argumentation to improve its positive perception or to increase the chances of fallacy identification?
- Is there any pre-validation/verification of the model that could have been done to improve the results?

# Sources (Online)

---

MIT Course on Fairness

MIT RES.EC-001 Exploring Fairness in Machine Learning, Spring 2020

<https://ocw.mit.edu/RES-EC-001S20>

USAID – United States Agency for International Development  
Artificial Intelligence and Machine Learning

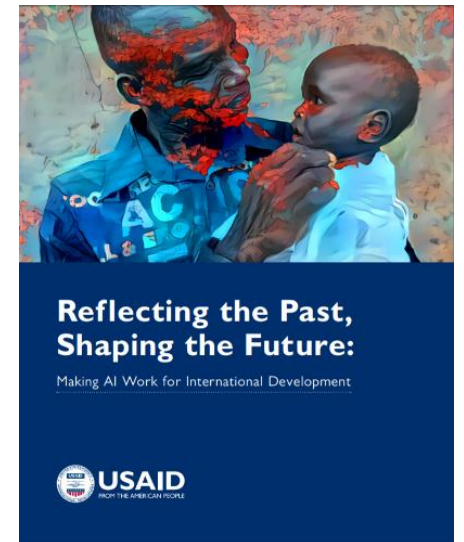
<https://www.usaid.gov/digital-development/artificial-intelligence>

Reflecting the Past, Shaping the Future: Making AI Work for International Development

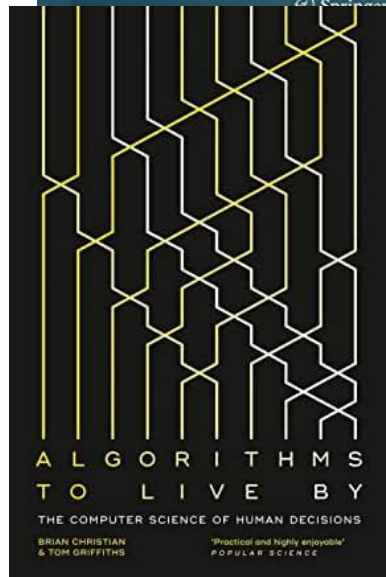
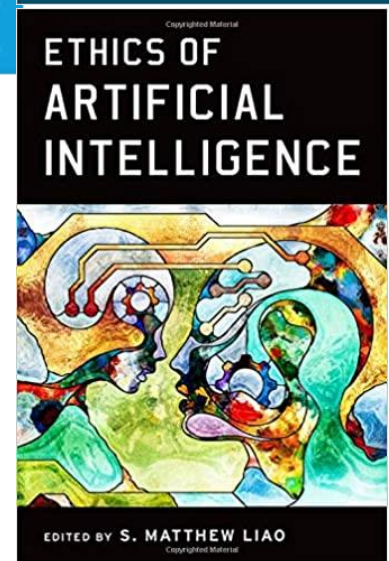
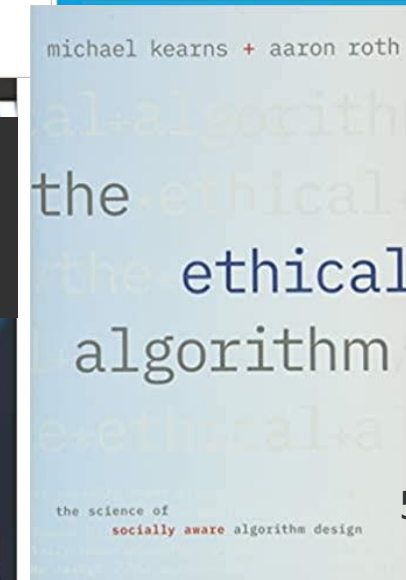
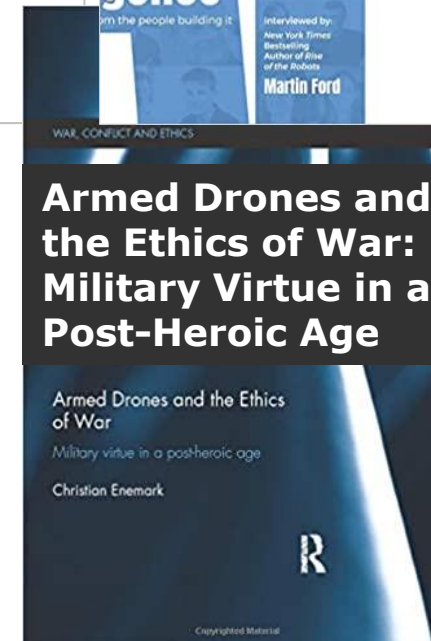
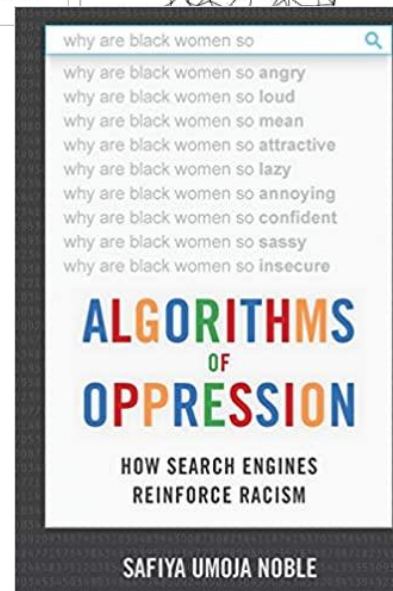
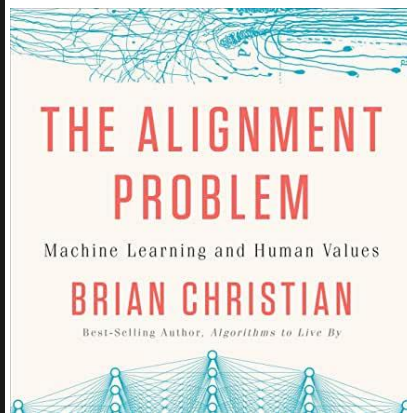
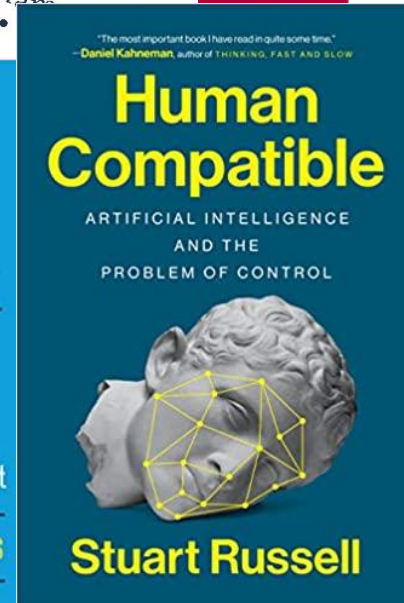
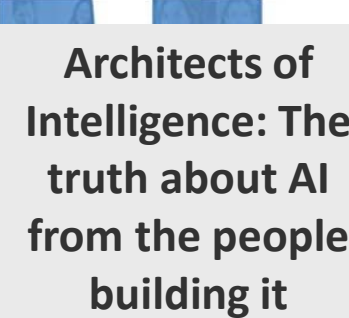
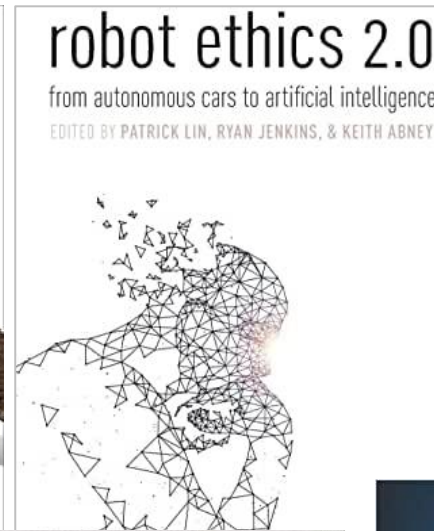
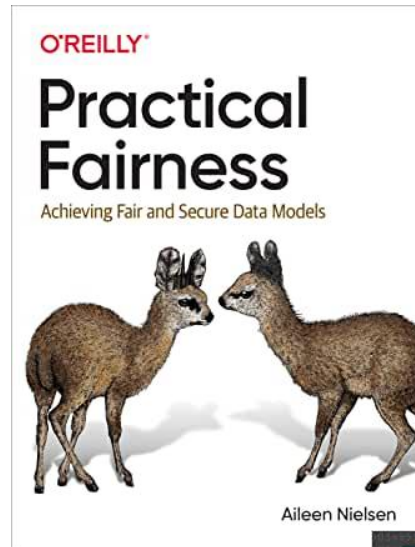
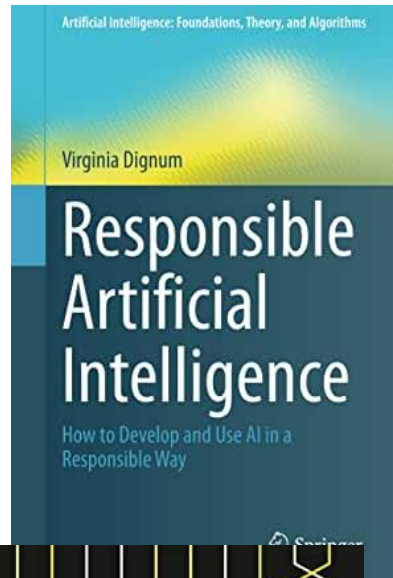
[https://www.usaid.gov/sites/default/files/documents/15396/AI\\_ExecutiveSummary-Digital.pdf](https://www.usaid.gov/sites/default/files/documents/15396/AI_ExecutiveSummary-Digital.pdf)

Discussion Group on Human-Centered AI (HCAI)

<https://groups.google.com/g/human-centered-ai>



# Sources (Books that we have been reading)





# How Good Models Can go Wrong [USAID 2020]

---

1. **Hard-Coded Inequities:** when real-world biases are present in training data, the resulting model will include the same biases.
2. **Opaque Models:** technical choices that favor prediction power can reduce model interpretability
3. **Misplaced Trust:** Overconfidence on models might lead to confirmation bias (only trust when the model agrees with one's beliefs)
4. **Wrong tool for the Job:** Models might not be aligned with decision-making needs.
5. **Active Misuse:** violations of privacy, spread of fake news, promotion of hate speech, prejudice messages

Can you think of concrete ethical dilemmas that entail trade-offs between these concerns versus the engineering principles of maximizing?

- efficiency (cost, reusability)
- efficacy (availability, speed)
- robustness (generalizability, portability)
- quality (accuracy)

**Relevance** - is it solving the correct problem?

**Representativeness** - is the data free of relevant bias?

**Value** – the new technology costs and risk justify its benefits

- Predictions are more accurate, and variances better explained than alternative methods?
- Are predictions "actionable" (causal/counterfactual models)?
- Are predictions timely and do they reach the right person?

**Explainability** - How well\* the process of building models and generating predictions communicated?

- \*well = transparent, tailored, understandable to guarantee trustworthiness 7

# Design Principles [USAID 2020][Gandhi 2020]

## Auditability

---



How well one can query and monitor the decision process that relies on the model?

Goals: ensure that decisions are fair (justice), unbiased (correct), and do not harm users (safety/secure/privacy compliant) Challenges: models are often harder to understand

Auditability may require specialized technical and legal infrastructure, because of complexity or restricted access to proprietary information.



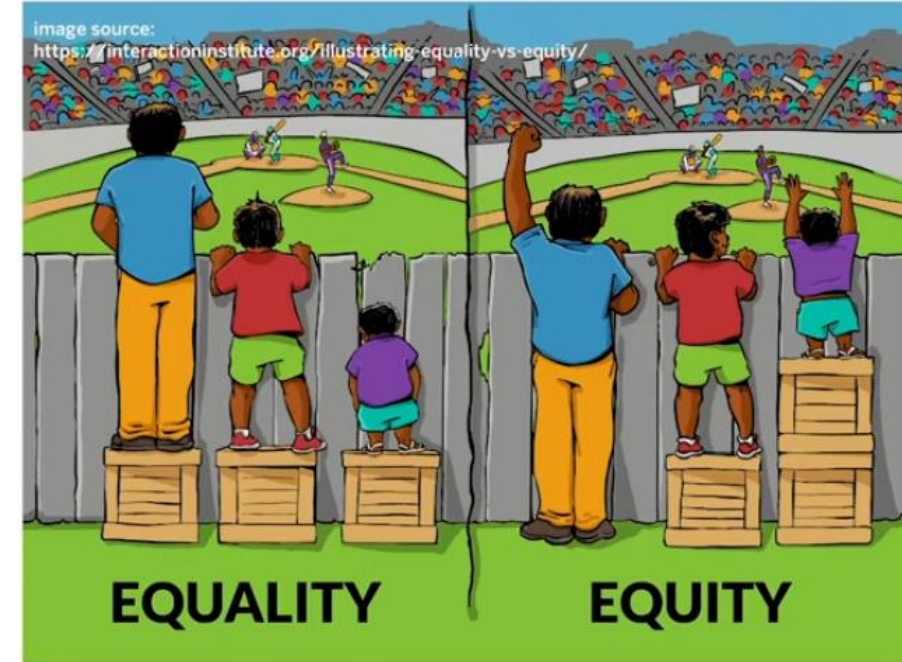
# Principles [USAID 2020][Gandhi 2020]

## Equity

How well the model prevents disproportionate benefit or harm to certain individuals or groups.

Does the model produce worse predictions for certain groups, e.g., face recognition failing depending on skin color?

How to measure it? Unexplained differences in the rates of false positives or false negatives.



# **Fairness Criteria** [USAID 2020][Gandhi 2020] **Accountability/Responsibility**

---



How well one can attribute credit or blame for decisions made by the model?

Challenge: many areas do not yet have a legal framework, for instance, like medical diagnosis and aviation have.

However, does a risk of not taking a decision because of legal support outweighs the consequences of inaction?

# Protected Attributes by Law (in some countries)

---

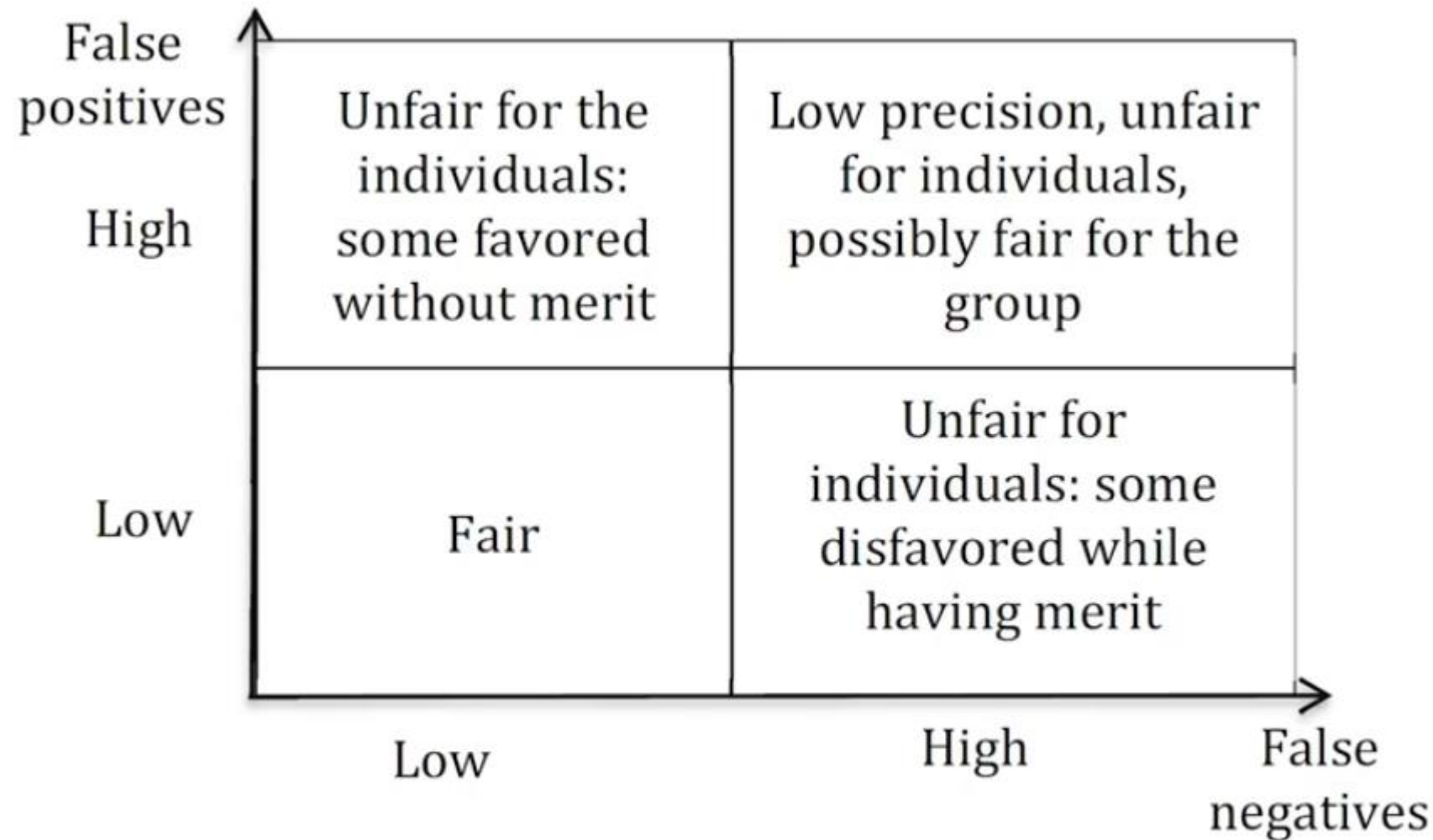
- age
- race
- skin color
- eye color
- height
- weight
- medical condition
- religion
- cultural heritage
- nationality
- citizenship
- SOGIE - sexual orientation, gender identity & expression
- marital status
- family size (has children or not)
- educational background
- income
- socioeconomic background
- etc.

Can you think of concrete ethical dilemmas that entail trade-offs between access to these attributes in order to engineer effective AI-systems?

# Confusion Matrix

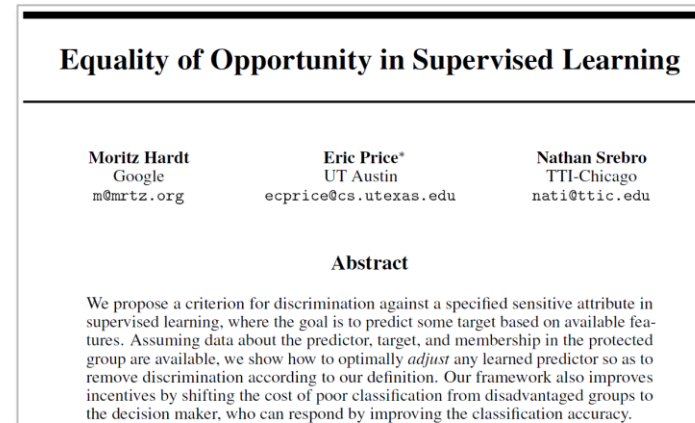
		Prediction	
		Negative	Positive
Actual	Negative	<b>True Negatives</b>	<b>False Positives</b>
	Positive	<b>False Negatives</b>	<b>True Positives</b>

# Fairness at the Individual Level vs Group Level



# Fairness criteria [Hardt et al. 2016][Zafar et al 2015]

- Fairness through unawareness
- demographic parity
- equalized odds
- equalized opportunity
- Individual fairness



[Hardt et al. 2016]



[Kusner & Loftus]

[Hardt et al. 2016] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 3315-3323.



# Fairness through awareness

---

**Definition** [Grgic-Hlaca *et al.* 2016]: remove any data that are considered prima facie to be unfair.

**Example** – Take an algorithm used by judges making parole decisions, fairness through unawareness could dictate that data on ethnic origin should be removed when training this algorithm, whereas data on the number of previous offences can be used.

**Caveats** [Wilford & Khairalla 2019]:

- data are usually and inevitably biased, e.g., number of previous offences can bear the stamp of historical racial bias in policing, as can the use of plea bargaining (pleading guilty being more likely to reduce a sentence than arguing innocence)
- Impossible trade-off: either remove all data or keep biased data.

# Demographic Parity

**Definition** [Hardt et al. 2016]: Demographic parity requires that a decision—such as accepting or denying a loan application—be independent of the protected attribute.

- e.g., accepting or denying a loan application should be independent of the protected attribute.

In the case of a binary decision  $\hat{Y} \in \{0,1\}$  and a binary protected attribute  $A \in \{0,1\}$ , this constraint can be formalized by asking that

$$\Pr\{\hat{Y} = 1 \mid A = 0\} = \Pr\{\hat{Y} = 1 \mid A = 1\}$$

**Caveats** [Dwork et al. 2012]:

- 1- It doesn't ensure fairness
- 2- It reduces utility of decisions that could be predicted with protected attribute

[Hardt et al. 2016] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 3315-3323.

[Dwork et al. 2012] Dwork, C., et al., (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

# Equalized Odds

**Definition** [Hardt et al. 2016]: A predictor  $\hat{Y}$  satisfies **equalized odds** with respect to protected attribute  $A$  and outcome  $Y$ , if  $\hat{Y}$  and  $A$  are independent conditional on  $Y$ .

$$\Pr\{\hat{Y} = 1 \mid A = 0, Y = y\} = \Pr\{\hat{Y} = 1 \mid A = 1, Y = y\}, \quad y \in \{0, 1\}$$

- Unlike demographic parity, equalized odds allows  $\hat{Y}$  to depend on  $A$  but only through the target variable  $Y$ . As such, the definition encourages the use of features that allow to directly predict  $Y$ , but prohibits abusing  $A$  as a proxy for  $Y$ .
- For the outcome  $y = 1$ , the constraint requires that  $\hat{Y}$  has equal true positive rates across the two demographics  $A = 0$  and  $A = 1$ . For  $y = 0$ , the constraint equalizes false positive rates.

**Caveat** - equalized odds enforces that the accuracy is equally high in all demographics, punishing models that perform well only on the majority.

# Equal Opportunity = giving the same beneficial predictions to individuals in each group

**Definition** [Hardt et al. 2016]: A binary predictor  $\hat{Y}$  satisfies **equal opportunity** with respect to  $A$  and  $Y$  if

$$\Pr\{\hat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\hat{Y} = 1 \mid A = 1, Y = 1\}$$

- It allows for stronger utility as shown by experiments in [Hardt et al. 2016].

## Caveats –

1. Equal opportunity is a weaker notion of non-discrimination
2. Typical societal unfairness is not captured by equality of opportunity.

# Individual Fairness

**Definition** [Dwork et al. 2012]: : similar individuals should get similar predictions.

**Example** - If two people are alike except for their sexual orientation, say, an algorithm that displays job advertisements should display the same jobs to both

**Caveats** [Dwork et al. 2012]:

Relies on a difficult concept of similarity

In this example, training data will probably have been distorted by the fact that one in five individuals from sexual or gender minorities report discrimination against them in hiring, promotions and pay [Pizer et al., 2011]

# Recent discussion on criteria for fairness

Build models that identify and mitigate the causes of discrimination.



A migrant farm worker has her fingerprints scanned so that she can register for a national identity card in India.

## The long road to fairer algorithms

Matt J. Kusner & Joshua R. Loftus

Build models that identify and mitigate the causes of discrimination.

relationships in data, we need models that capture or account for the causal pathways that give rise to them. Here we outline what is required to build models that would allow

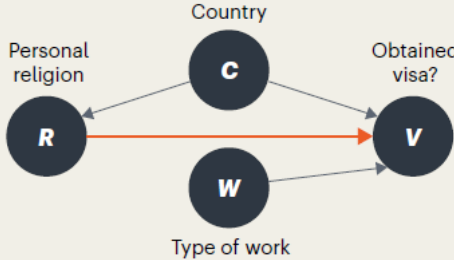
visa (see ‘Three causal tests’, part 1). This model says that the country of origin directly influences a person’s religion and whether they obtain a visa; so, too, do religion and type of work. Having a causal model allows us to address questions related to ethics, such as does religion influence the visa process? But because many different causal models could have led to a particular observed data set, it is not generally possible to identify the right causal model from that data set alone<sup>3</sup>. For example, without any extra assumptions, data generated from the causal graph described here could seem identical to those from a graph in which religion is no longer

[Kusner & Loftus] Kusner, M. & Loftus, J., (2020), The long road to fairer algorithms, Nature.

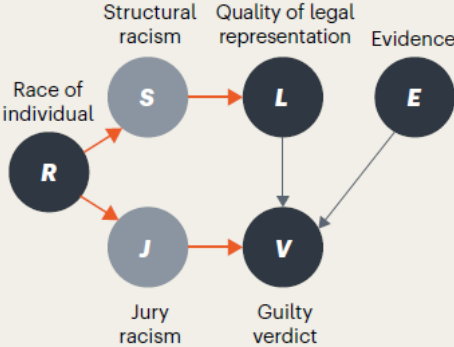
### THREE CAUSAL TESTS

Algorithmic fairness can be examined in different ways.

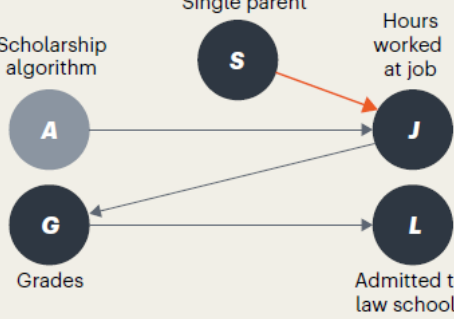
1. Counterfactuals



2. Sensitivity



3. Impacts



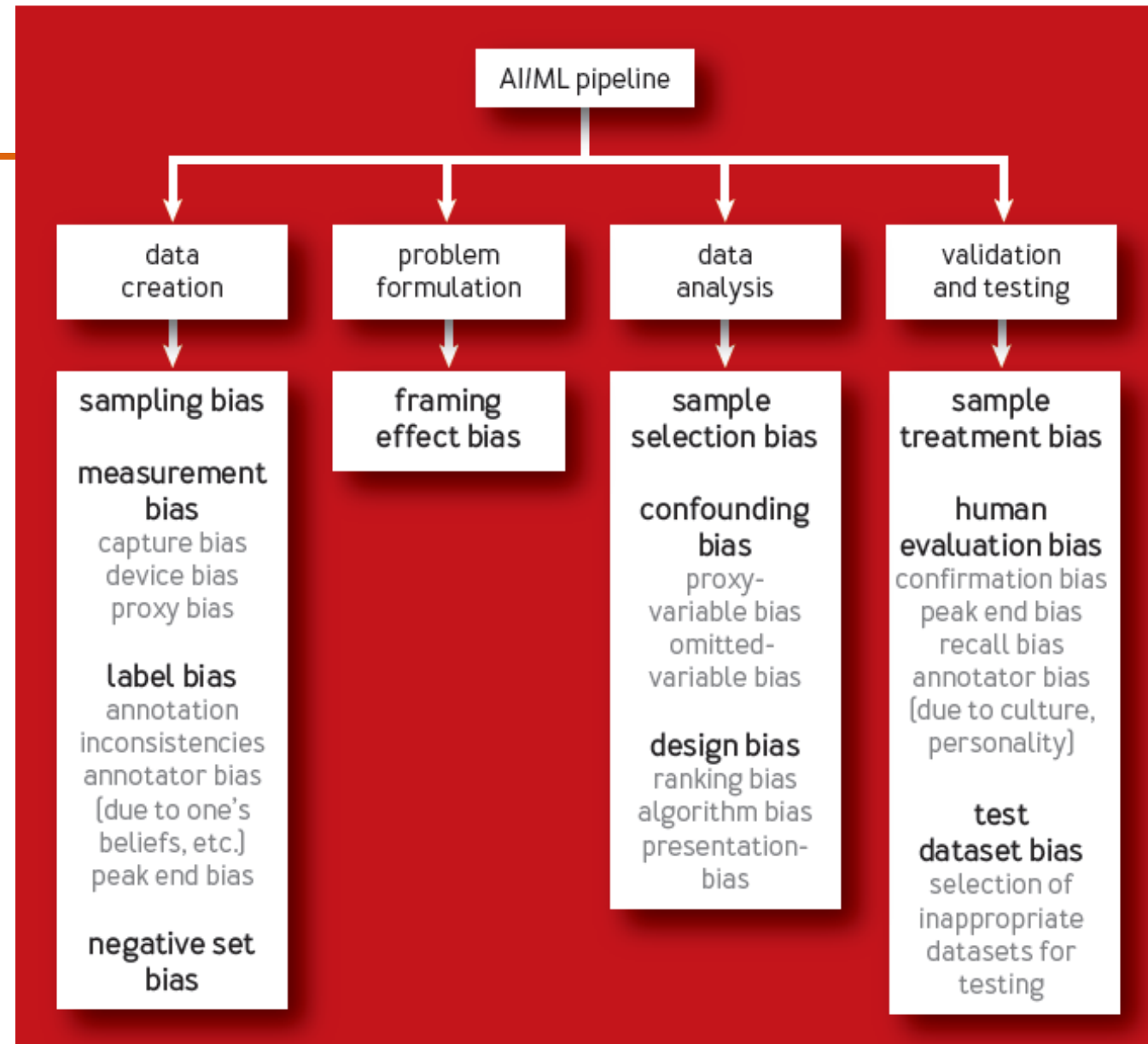


# Overview of Design Space of Solutions

# Taxonomy of Biases

- There are many reasons for an engineer to have a wrong model of the world (figure-1)
- These biases also impact users in very diverse ways.
- I am more interested on bias sample selection bias and confounding bias (under the data analysis)
- Before we delve into these bias, we need to answer the question, **why many times simply getting more data does not solve the bias problem?**
- The Reason: the bias-variance trade-off

FIGURE 1: TAXONOMY OF BIAS TYPES ALONG THE AI PIPELINE



# How do we currently think about robustness?

-> Bias-Variance Trade-off

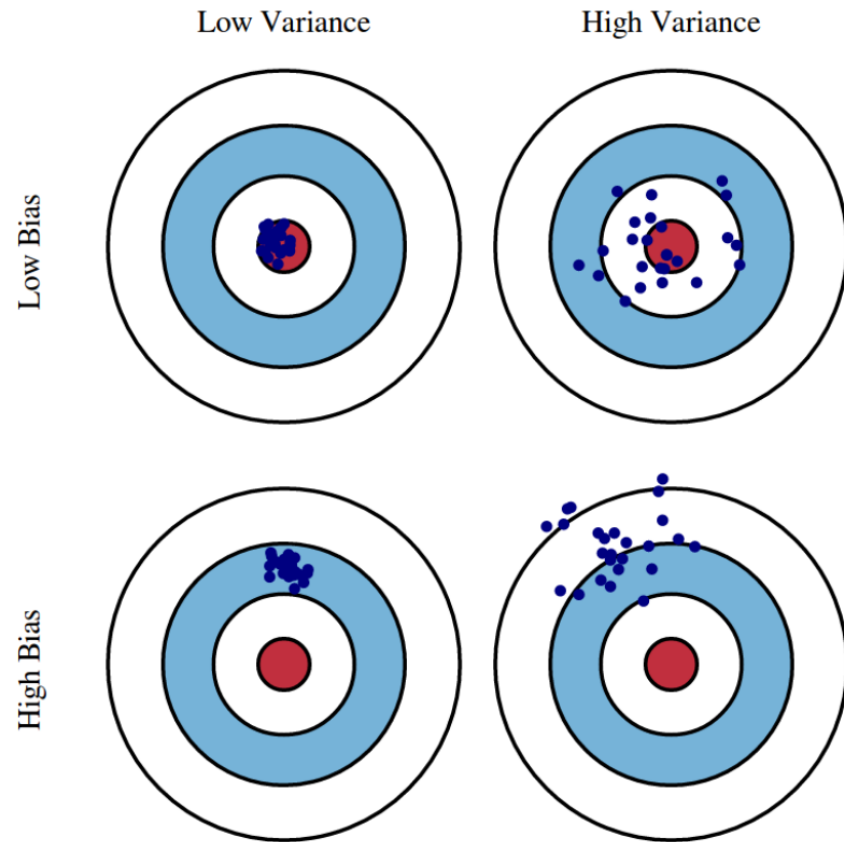


Fig 1: Graphical illustration of bias and variance.

Source: <http://scott.fortmann-roe.com/docs/BiasVariance.html>

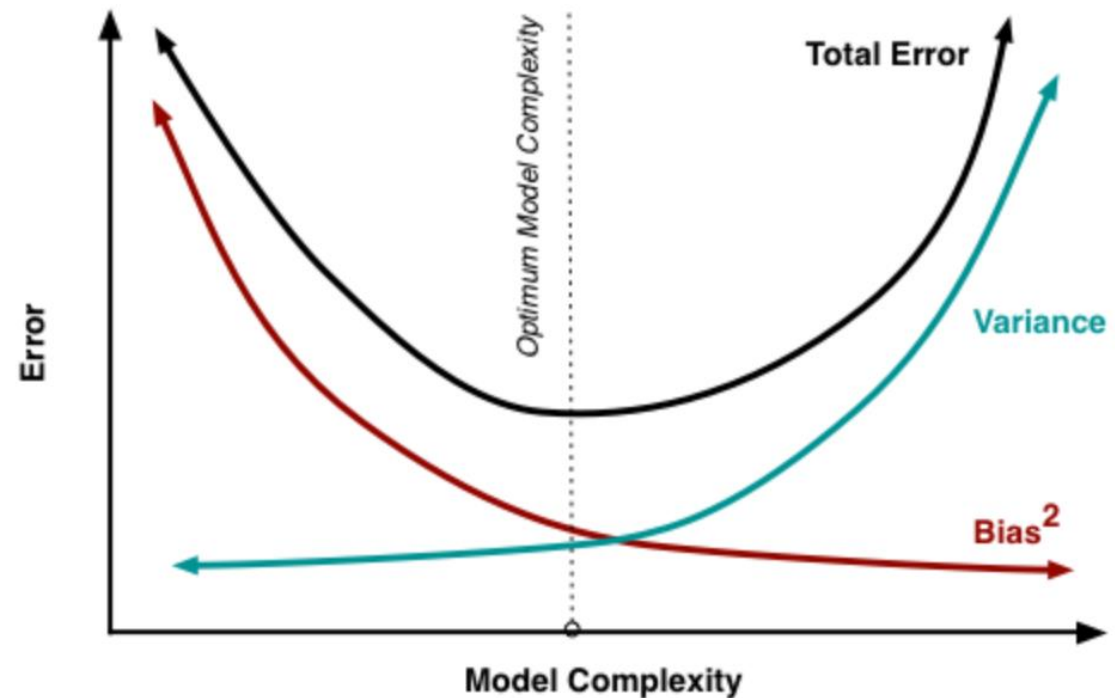


Fig 2: The variation of Bias and Variance with the model complexity. This is similar to the concept of overfitting and underfitting. More complex models overfit while the simplest models underfit.

Source: <http://scott.fortmann-roe.com/docs/BiasVariance.html>

# Implications to predictive models

---

**Goal:** Generalize data associations as predictive patterns

**Assumptions:** good data and observable patterns

**Reality:** sparse data and hidden states

- Sparse data (Essential limitation, cannot eliminate with better prediction models)
- Latent patterns (Accidental, can eliminate with better models)
  - Source – Misspecification

**Not enough data or bad tuning** of a model can make the concept drift more severe, as models might present strong bias (insensitive to crucial features) or high variance (too sensitive to noise).

- Under-specification (leads to bias-underfitting)
- Over-specification (leads to variance-overfitting)

# Adversarial Changes in the Environment Sources of Sparsity and Unobservability

---

Changes in the Data Generation Process:

- Covariate Shift (change in data distribution)
- Domain Shift (change in the action-state space)
- Concept Drift (change in the associations)

These changes are independent of the model, but the model might make the problem worse.

**Goal:** A robust model should have structures and conditions in place to mitigate the effect of these changes on the performance of the model.

**Plausible Changes -> Sparsity + Observability -> Model performance**

# Mitigation of Ethical Failures / Dilemmas

## Data-Centric vs Systems-Centric

---

**Data-Centric:** Which data problems (data privacy violations, biases, etc.) are ethical dilemmas or failures in machine learning models? Mitigation might involve, pre-process data, obtain better data, augment data, or use more robust statistical methods (e.g., less prone to overfitting).

**Systems-Centric:** Which levels of autonomy contribute to ameliorate or degrade the ability of a system to handle ethical dilemmas?

- **Design Aspects** - feedback loops, cross-cutting concerns (monitoring, exception handling, failure propagation), decision-making mechanisms (agents, controllers), and the corresponding actuators.
- **General Goal** - How AI-Systems can self-adapt to cope with adversarial changes in the environment that impose ethical failures / dilemmas

**Example:** to redesign an avionics system to comply to a more appropriate set of ethical requirements one needs to go beyond better prediction models. One needs to understand which aspects of the system contribute to ethical failures and how these can be mitigated. Noting that mitigation actions might involve new ethical choices.



# Adversarial Fragilities

## Online (Continual) Learning

Hidden confounders + Selection Bias

Simpson's and Berkson's paradoxes

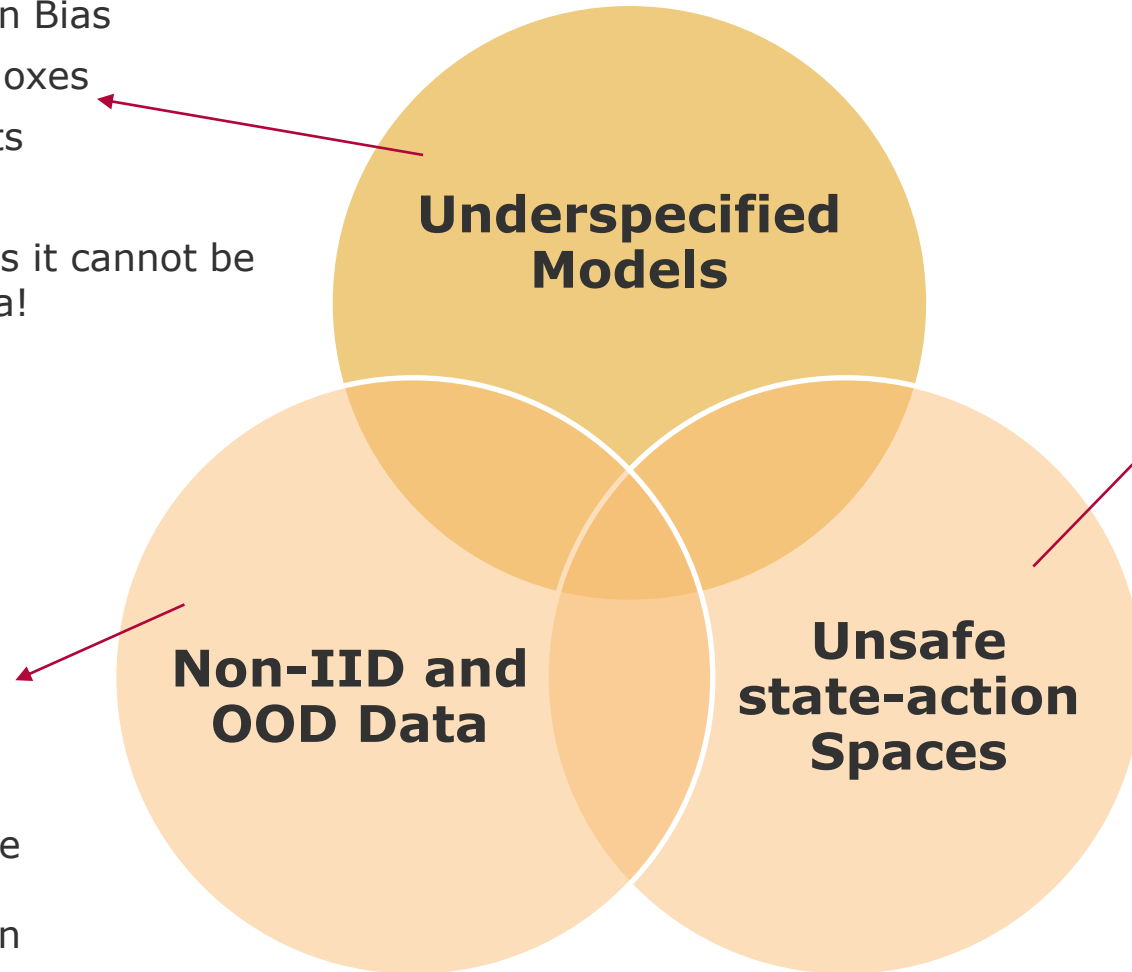
Shortcut learning in Neural Nets

This goes beyond overfitting, as it cannot be solved with more or better data!

Real-world is non-stationary

Predictions affects the data generation process

Modeling better recommender systems is not enough, because uncertainty grows wildly when extrapolating out-of-distribution



Wrong predictions can spur unsafe actions that can lead to unsafe states.

Sensitivity analysis and testing on hold-out-sets are ad hoc approaches cannot guarantee safety.

"Program testing can be used to show the presence of bugs, but never to show their absence!" — Edsger W. Dijkstra

# Adversarial Fragilities - Solutions

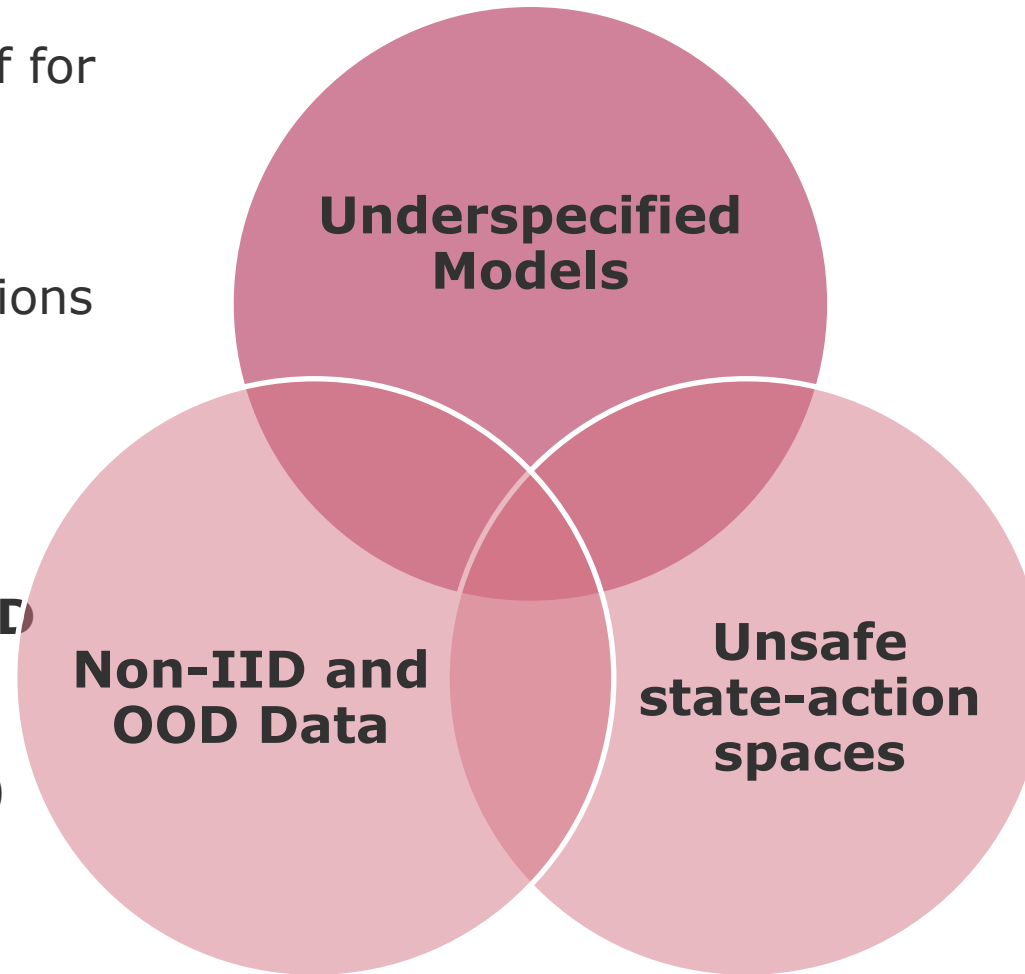
## Online (Continual) Learning

### Pre-trained models:

- Optimal Bias-Trade-off for fine-tuning
- Rashomon sets
- Domain-Adaptation
- Invariant Representations

### Generative models of OOD

- Sampling-based (importance sampling)
- Feature-based (new task)
- Intervention-based (control)



### Safe learning in production:

- Specify safety (domain-knowledge, Constrained MDP, Shielding)
- Learn to be safe (sandbox/digital-twin)

# Connections between Fairness and Robustness

---

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). **Mitigating unwanted biases with adversarial learning.** In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). **Avoiding discrimination through causal reasoning.** arXiv preprint arXiv:1706.02744.

## Suggested questions

- How do failures in satisfying fairness requirements lead to lack of robustness?
- What type of robustness engineering methods can be used to reify fairness requirements?
- How fairness can be achieved by means of Learning to be Safe, Fail-Safe Resilient, Recoverable mechanisms?
- ?

# Warm-up task: Start thinking about ethical dilemmas, cases and solutions

---



- Choose a domain and an ethical dilemma of your interest
- Research two definitions for the same ethical principle
- Research one fail case with a corresponding mitigation (data and system-centric)
- Describe briefly (~4 lines) and write down your opinion (~4 lines)
- Share it on Slack by Monday Evening (channel #writting)