Winter Term 21/22

Artificial Intelligence, Ethics & Engineering

Lecture-5: Experimental Philosophy

Prof. Dr. Holger Giese (holger.giese@hpi.uni-potsdam.de)

Christian Medeiros Adriano (christian.adriano@hpi.de) - **"Chris"**

Christian Zöllner (Christian.zoellner@hpi.de)

The role of the philosopher’s intuition

- “… A familiar form of analytic technique consists of proposing **necessary** and **sufficient** conditions for the truth of a schematized form of sentence containing a term under investigation.
- Philosophers ask what it is for A to know that p, or under what conditions it is true that e causes e’, or what makes it true that “A did a intentionally.”
- Then they run through various previously pro-posed necessary and sufficient conditions, rejecting in turn any that are open to **counterexample**.
- “a counterexample is a situation in which the proposed necessary and sufficient conditions obtain but the philosopher is disinclined to say, or is inclined to deny, that A knows that p or that e causes e’ and so on. In these cases, the inclination or disinclination is driven by the philosopher’s own (semantic) **intuitions**”

Sorell, T., 2917, Scientism (and Other Problems) in Experimental Philosophy

Strides of Science on Philosophical Questions

- Study morality with methods 'suitable to the investigation of natural facts' (Jesse Prinz)
- Results of fMRI scans refute deontologism (Joshua D. Greene, Peter Singer)
- Experimental psychology refutes moral intuitionism (Walter Sinnott-Armstrong) and virtue ethics (Gilbert Harman, John Doris)
- David Hume moral sentiment derives from oxytocin, 'the moral molecule' (Paul J. Zak, Patricia Churchland)
- Primatologists, evolutionary biologists, psychologists, and neuroscientists advocate the replacement of armchair ethics with an empirical
  - 'science of morality' (Sam Harris)
  - 'the science of good and evil' (Michael Shermer),
  - 'the science of our moral dilemmas' (Michael S. Gazzaniga)

The Manifesto [Knobe & Nichols 2007]

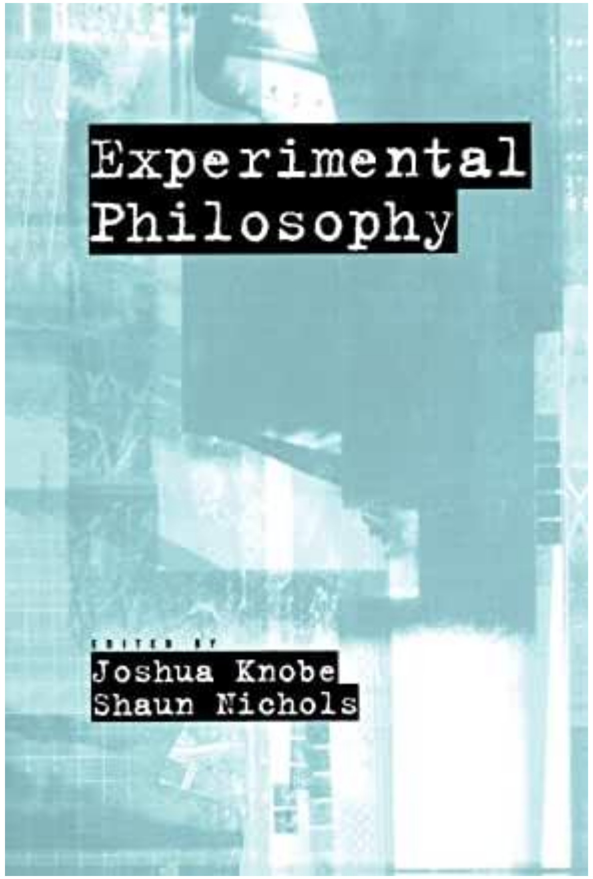
It used to be a commonplace that the discipline of philosophy was deeply concerned with **questions about the human condition**. Philosophers thought about human beings and how their minds worked. They took an interest in reason and passion, culture and innate ideas, the origins of people’s moral and religious beliefs. On this traditional conception, **it wasn’t particularly important to keep philosophy clearly distinct from psychology, history, or political science**. Philosophers were concerned, in a very general way, with questions about **how everything fit together.**

… experimental philosophy seeks a return to this traditional vision.. will involve us in **the study of phenomena that are messy, contingent, and highly variable across times and places…**

[However] Unlike the philosophers of centuries past, we think that a critical method for figuring out how **human beings think** is to go out and actually run systematic empirical studies. Hence, experimental philosophers proceed by **conducting experimental investigations** of the **psychological processes** underlying people’s **intuitions** about central **philosophical issue**,

Knobe & Nichols 2007, Experimental Philosophy

Sources (Books that we have been reading)

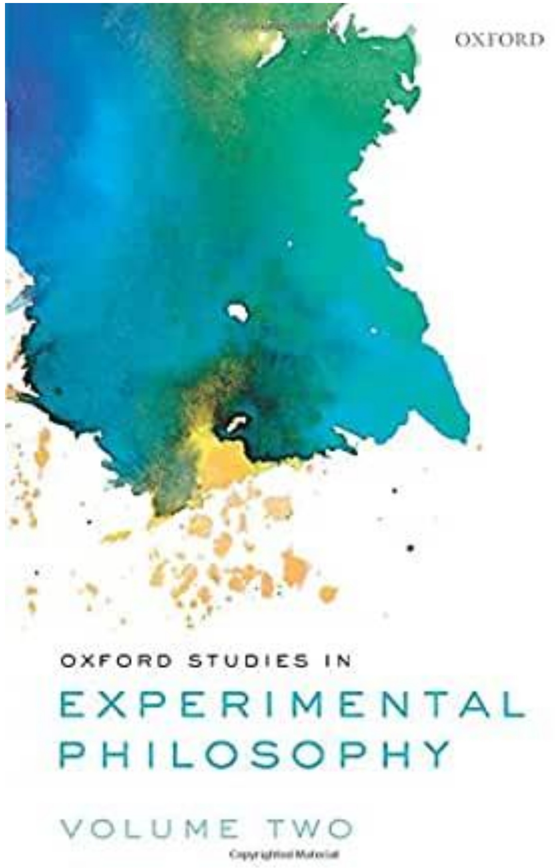


Experimental Philosophy

EDITED BY

Joshua Knobe

Shaun Nichols



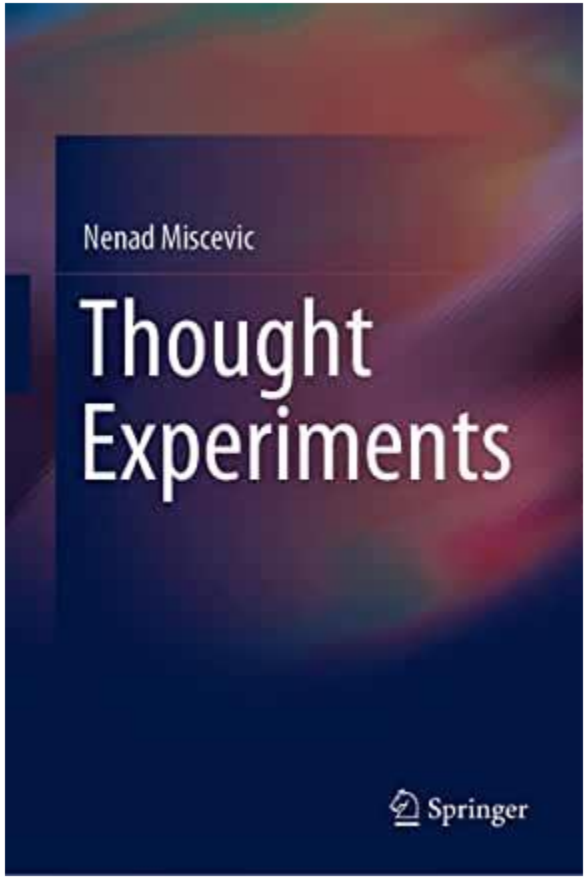
OXFORD

OXFORD STUDIES IN

EXPERIMENTAL PHILOSOPHY

VOLUME TWO

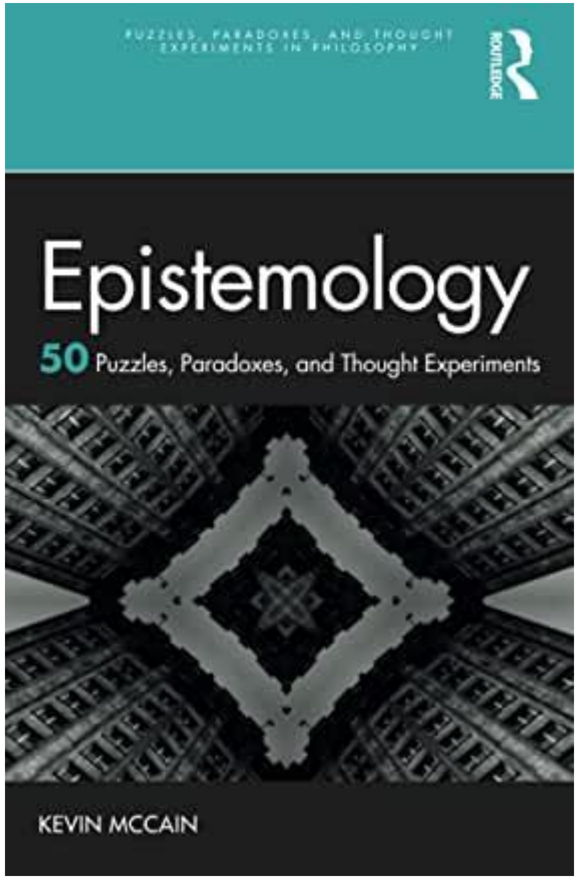
Copyrighted Material



Nenad Miscevic

Thought Experiments

Springer



PUZZLES, PARADOXES, AND THOUGHT EXPERIMENTS IN PHILOSOPHY

ROUTLEDGE

Epistemology

50 Puzzles, Paradoxes, and Thought Experiments

KEVIN MCCAIN

Sources (Online)

A Characteristic Difference- When Experimental Philosophy Meets Psychology - A Conversation between Joshua Knobe, Daniel Kahneman https://www.edge.org/conversation/joshua_knobe-daniel_kahneman-a-characteristic-difference

The New Science of Morality, Part 8

Joshua Knobe

https://www.edge.org/conversation/joshua_knobe-the-new-science-of-morality-part-8

Sytsma, J., & Livengood, J. (2019). **On experimental philosophy and the history of philosophy: a reply to Sorell**. *British Journal for the History of Philosophy*, *27*(3), 635-647.

Tom Sorell (2018) **Experimental philosophy and the history of philosophy**, British Journal for the History of Philosophy, 26:5, 829-849

The Gettier Problems [Gettier 1963]

The Nature of knowledge: When do we know that we know something?

This is important because if we do not fully understand it, we could never fully understand ourselves either.

→We need a rigorous way to define what is it to know something.

→Traditional definition: knowledge of a proposition is a justified true belief in that proposition.

Propositional knowledge is usually stated in the following format "knowledge that **p**," for instance, that the "knowledge that **Swans have wings**". "p" is a truth or fact – a knowledge of how the world is, regardless of how we describe it by an arbitrary instance of "**p**".

However, Gettier showed cases in which a **belief** can be both true and well supported by evidence but fails to be knowledge (for most epistemologists).

Justified-True-Belief Analysis of Knowledge [iep 2021]

Belief: A person *believes* that p. This can be expressed in words (or not) and could be a more or less confident belief. The only thing necessary to make it a belief is that it exists and to be true and justified.

Truth: belief that p needs to be *true*. If the belief is incorrect (false), then, regardless of any quality of goodness or utility value, the believe cannot be considered knowledge.

Justified: belief that p needs to be well *supported* by good evidence or reasoning, or rational justification. Otherwise, the belief, even being true, it will be only a guess. In other words, although it could be correct, belief it is not knowledge without proper justification.

However, Gettier showed cases in which a **belief** can be both true and well supported by evidence but fails to be knowledge (for most epistemologists).

The Gettier Case

1. Smith and Jones have applied for a particular job.
2. Smith has been told by the company president that Jones will win the job.
3. Smith knows from his observational evidence of there being ten coins in Jones's pocket, which he.
4. Smith infers that whoever will get the job has ten coins in their pocket (call it belief **b**) Notice that Smith is not thereby guessing. The belief if true and reasonably justified.
5. However, belief b is true in a different way - it is *Smith* who will get the job, and Smith *himself* also has ten coins in his pocket.
6. These two facts combine to make his belief b true. Nevertheless, neither of those facts is something that, on its own, was known by Smith.

Given this reasoning, is Smith's belief b, therefore not knowledge?

source: https://iep.utm.edu/gettier/

Other Gettier Cases

The Lucky disjunction

The Sheep in the field

The Pyromaniac

The Fake Barn

Still, in none of those cases (or relevantly similar ones), say almost all epistemologists, is the belief in question knowledge.

Attempted Solutions to Gettier Cases

Infallibility - allowing fallible justification is all that it would take to convert a true belief into knowledge. **Caveat** – how much infallibility?

Eliminate Luck – does not allow too much luck, otherwise, we would again have reached for the Infallibility Proposal. **Caveat** - How much luck is too much?

Eliminate False Evidence – no belief is knowledge if the person’s justificatory support for it includes something false. **Caveats** – (1) no false evidence can still be used to produce Gettier cases and (2) impossibility of complete absence of falsity in anyone’s belief

Attempted Solutions to Gettier Cases

Eliminate Defeat - what is not included in the person’s evidence: specifically, some notable truth or fact is absent from her evidence. what is needed in knowing that p is an absence from the inquirer’s context of any **defeaters** of her evidence for p.

A **defeater** is a particular fact or truth t defeats a body of justification j (as support for a belief that p) if adding t to j, thereby producing a new body of justification j*, would seriously weaken the justificatory support being provided for that belief that p — so much so that j* does not provide strong enough support to make even the true belief that p knowledge.

Insofar as one wishes to have beliefs which are knowledge, one should only have beliefs which are supported by evidence that is not overlooking any facts or truths which — if left overlooked — function as defeaters of whatever support is being provided by that evidence for those beliefs.

Caveat – how strong a defeater should be that it should not be overlooked?

Attempted Solutions to Gettier Cases

Eliminate Inappropriate Causality – belief should be caused — generated, brought about — in a normal way for it to be knowledge.

Caveats –applying only to empirical or a posteriori knowledge, knowledge of the observable world and difficulties with indirect causality and effects of absence of action (the gardener’s failure to water the plants causes their death).

In summary

When epistemologists claim to have a strong intuition that knowledge is missing from Gettier cases, they take themselves to be **representative of people in general** (specifically, in how they use the word "knowledge" and its cognates such as "know," knower," and the like).

That intuition is therefore taken to reflect **how "we" — people in general — conceive of knowledge**. It is thereby assumed to be an accurate indicator of pertinent details of the concept of knowledge — which is to say, "our" concept of knowledge.

So:

- What evidence should epistemologists consult as they strive to learn the nature of knowledge?
- Should they be perusing intuitions? If so, whose? Their own?
- How should competing intuitions be assessed?
- And how strongly should favored intuitions be relied upon anyway?
- Are they to be decisive? Are they at least powerful?
- Or are they no more than a starting-point for further debate — a provider, not an adjudicator, of relevant ideas?

Experimental Philosophy
The four projects [carneades.org]

Psychology Project: how people reason matters.

- Goal: study the ways in which regular people think about philosophical questions and underlying theories.
- Methods: Psychology methods

Verification Project: how philosophers disagree on each others' intuitions matters.

- Goal: discover which propositions or positions normal people find intuitive
- Methods: Sociology methods

Sources Project: how intuitions arise matters

- Goal: identify intuitions (i.e., w.r.t. moral questions) that arise through inappropriate means
- Methods: use of fMRIs to distinguish between neurological processes

Variation Project: reliability of intuitions across context and culture matters

- Goal: identify intuitions that vary across context (which might invalidate them)
- Methods: Anthropology methods

Next Task

Part-1 Study the other four Gettier cases

For tomorrow, Nov. 17

- The Lucky disjunction
- The Sheep in the field
- The Pyromaniac
- The Fake Barn

For Monday, Nov. 23

Part-2 Think of ways to inform our reasoning about a particular ethical dilemma, for instance, by obtaining answers to the following questions:

- What are other popular reasoning of the particular dilemma?
- How intuitive (easy to understand) people find certain reasonings?
- How variations in phrasing affect people judgments?
- When do people distinguish between the notion of knowing something versus just believing in something?

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, mis-understandings, etc.
- 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?

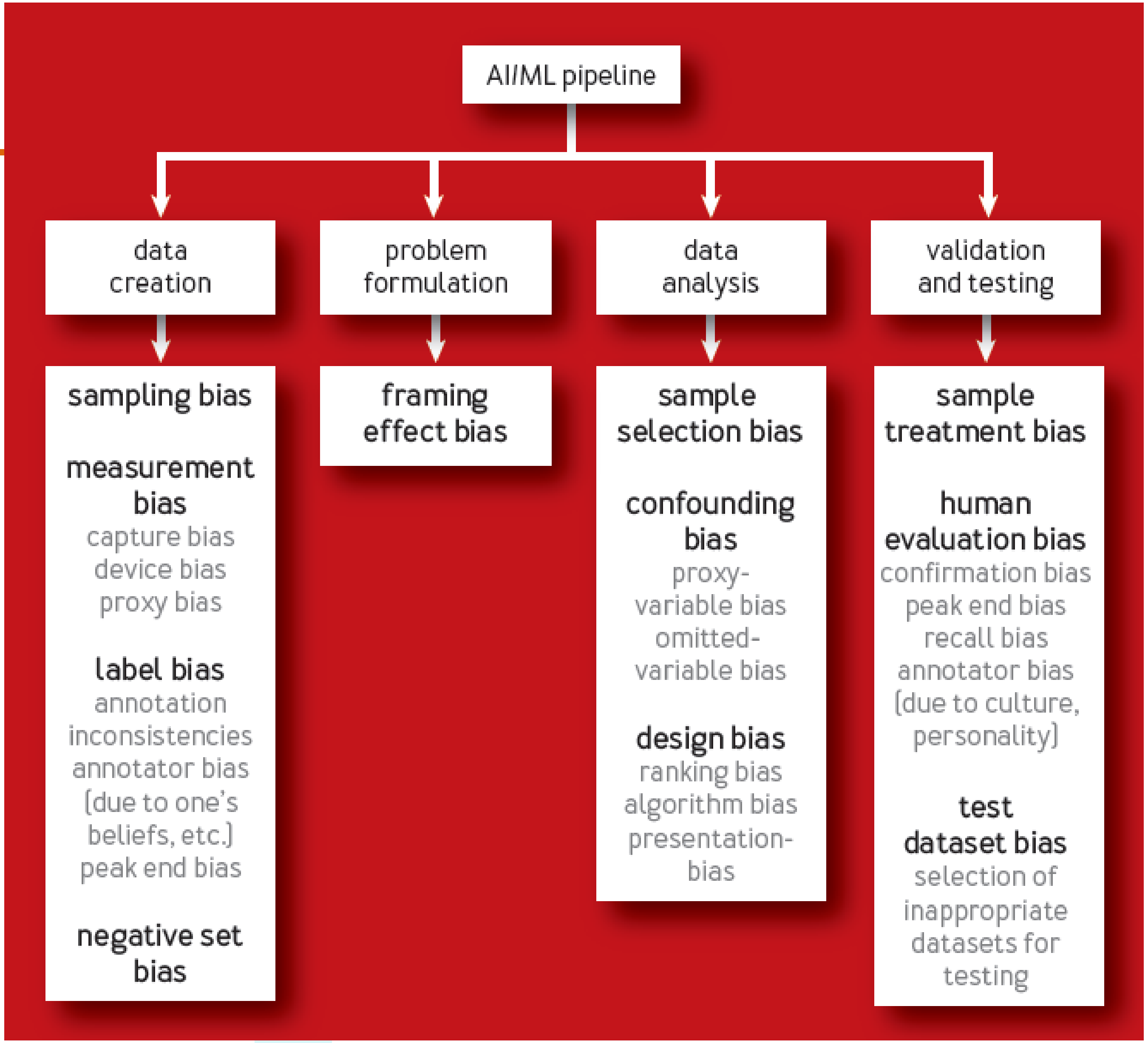
End

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



AI/ML pipeline

data creation

sampling bias

measurement bias
capture bias
device bias
proxy bias

label bias
annotation inconsistencies
annotator bias (due to one's beliefs, etc.)
peak end bias

negative set bias

problem formulation

framing effect bias

data analysis

sample selection bias

confounding bias
proxy-variable bias
omitted-variable bias

design bias
ranking bias
algorithm bias
presentation-bias

validation and testing

sample treatment bias

human evaluation bias
confirmation bias
peak end bias
recall bias
annotator bias (due to culture, personality)

test dataset bias
selection of inappropriate datasets for testing

[Srinivasan & Chander 2021]

How do we currently think about robustness? -> Bias-Variance Trade-off

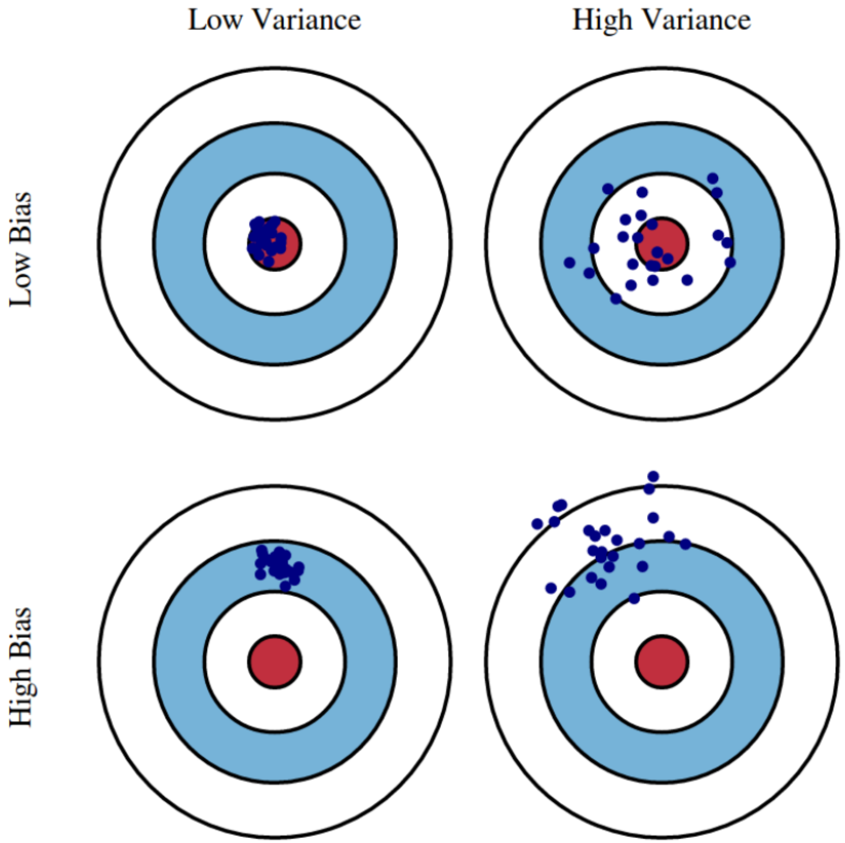


Fig 1: Graphical illustration of bias and variance.
Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

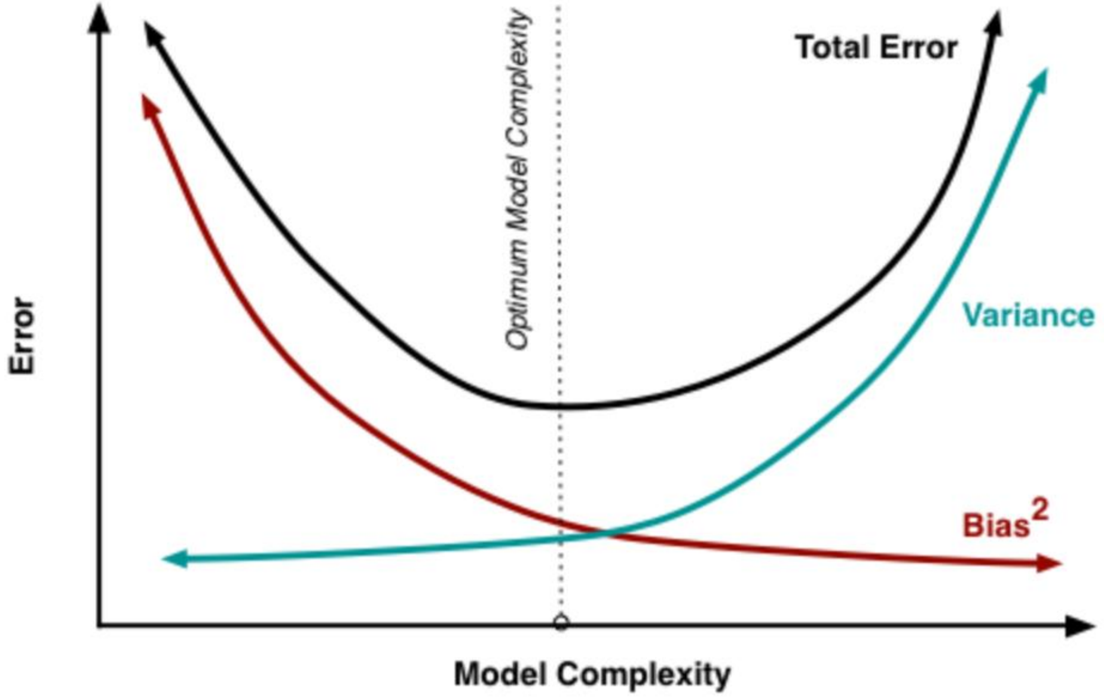


Fig 2: The variation of Bias and Variance with the model complexity. This is similar to the concept of overfitting and underfitting. More complex models overfit while the simplest models underfit.
Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

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 tunning 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

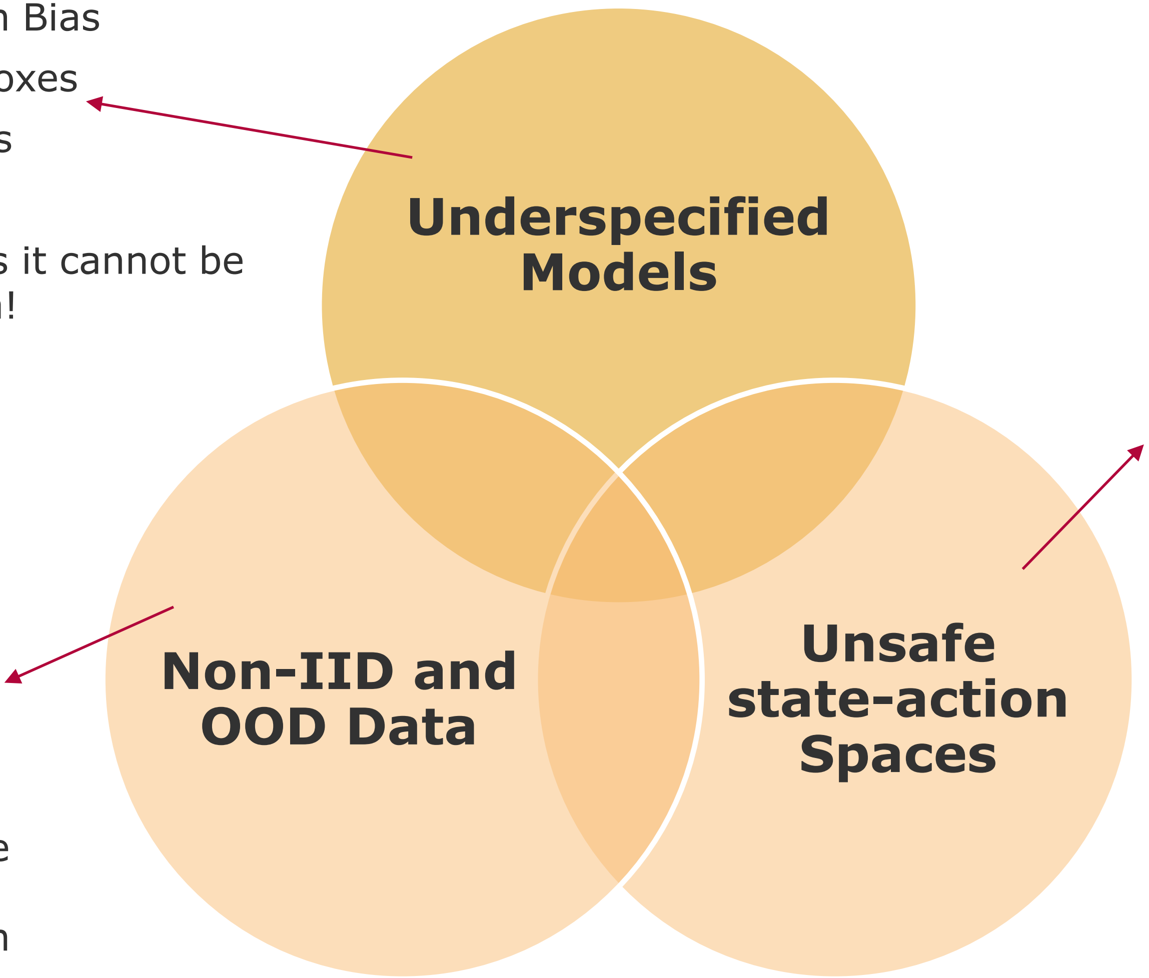
Online (Continual) Learning

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.

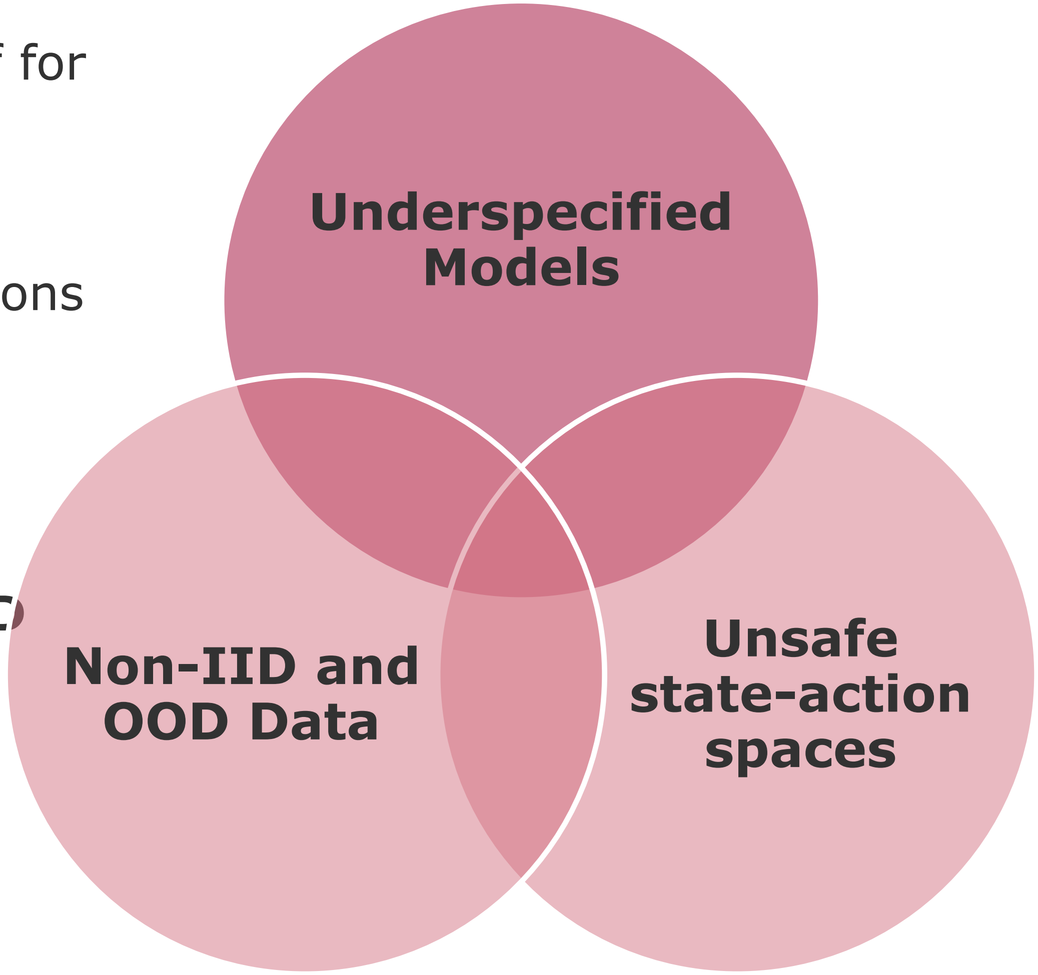
“Program testing can be used to show the presence of bugs, but never to show their absence!” — Edsger W. Dijkstra

Adversarial Fragilities - Solutions

Online (Continual) Learning

Pre-trained models:

- Optimal Bias-Trade-off for fine-tuning
- Rashamon 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)