Research statement: past work

Reasoning from examples is one of the most common methods in philosophy. A philosopher describes a case, asks what readers judge about it, and treats their judgements as evidence for or against a general philosophical hypothesis. My research asks how different forms of this method are supposed to work, what threatens them, and how they might do better by their own lights.

In this work, I draw on methods and inference-forms in formal epistemology, statistics, causal inference, metaethics and the philosophy of literature. For instance, certain philosophers who reason from examples model their reasoning—for better or for worse—on scientific experiment. An ethicist might compare two cases that differ in a single respect—an agent kills in one and lets die in the other, say—and attribute any difference in judgement to that respect, as an experimenter attributes a difference in 'outcome' to the 'treatment'. I take them at their word: *if* their reasoning is a kind of experiment, they are committed, by their own lights, to guarding against analogues of the errors that statistics and causal inference have well documented in inferences of that kind. I argue that this commitment is more demanding than is usually recognised.

My dissertation comprises four papers. Each concerns the choices involved in representing a situation, whether as a described case, or as a formal model such as a decision problem. For starters, a philosopher must decide *which cases* to consider. But she must also decide which *details* to specify as well as how to *refer* to the individuals involved described. The method leaves these choices open. They are not innocent: they can reverse the results of the reasoning. The papers ask when, why, and what to do about it.

My essay 'The granularity tradeoff' (currently at the 'revise and resubmit' stage at a major philosophy journal) asks how much detail a case should contain, and argues that no level of detail is free from danger. Sparse cases, of the kind philosophers usually use, leave readers to fill in what is unstated, and different readers are likely to fill it in differently. Detailed cases provoke more disagreement, and what is learned from them may not carry over to other situations. The choice between these risks is the counterpart of a tradeoff familiar from statistics; between models too simple to capture a pattern, and models fitted so closely to their data that they fail to generalise. Like that tradeoff, it can be managed but not escaped. I argue the tradeoff faces every form of reasoning from examples, within philosophy or outside it. Since the right level

of detail depends on the inquiry at hand, the paper recommends that philosophers say which risks their cases run and check whether their verdicts survive reasonable changes of detail.

Relatedly, my essay 'Reasoning from examples is threatened by Simpson's paradox', co-authored with Professor Richard Pettigrew, identifies one way in which *sparse* cases can mislead. In Simpson's paradox, a trend in a population reverses within each of its subgroups: a treatment can have a higher recovery rate than its rival among patients as a whole, and a lower one both among men and among women separately. We argue that judgements about cases can reverse in the same way. If, as is plausible, a reader's judgement about a sparse case draws on her beliefs about how the features of such situations tend to go together, it can reverse once a detail the case left open is filled in—*whichever* way it is filled in. We also argue that a verdict which survives the addition of detail may *still* be a poor guide to action, depending on causal facts that the case does not settle. The worry bears on influential arguments about the wrongness of killing and about our obligations to future people, and the paper identifies which details a philosopher must specify to guard against it.

Simpson's paradox is mentioned in the more general paper 'Analogues of fallacies in statistical reasoning afflict reasoning from examples in philosophy'. Philosophers who compare cases that differ in a single respect aim, like experimenters, to isolate the effect of one feature. I argue that this aim commits them, by their own lights, to guarding against analogues of seven well-known statistical errors, and I show how each can arise. The framework I develop in the dissertation is also constructive and optimistic. It explains, for instance, why Kripke's (1981) case of Gödel and Schmidt tells so strongly against the view known as Descriptivism.

The final paper comprising the dissertation, 'Model behaviour: solving Mahtani's puzzles without the tenet', carries these questions from described cases to formal models. Anna Mahtani has presented puzzles in which equally natural ways of applying Bayesian epistemology or decision theory to the same situation give conflicting results. I argue that these are problems of choosing between models. They can be solved by asking which model best represents the reasoner's situation, without the thesis about belief on which Mahtani's own explanation relies. I also argue that they share a source with a long-standing difficulty in the theory of probability, *the reference class problem*—the problem of deciding which group of similar cases should fix a probability. The paper ends with two constructive rules for representing a situation as a decision problem.

Formal models are prized for their precision. One might hope that they avoid the difficulties of informal cases. The dissertation shows that they face the same choices about how to represent a situation. It gives an account of what reasoning from examples can show and where it goes wrong. Every form of the method depends on choices about how to represent an amorphous situation, and the papers ask when those choices can be trusted. Statistics and causal inference serve throughout as tools of diagnosis. They also serve as a source of safeguards for those whose reasoning is

already modelled on science. My future research extends this work in three directions: to cases that are interpreted as wholes, often drawn from literature; to the failures that machine learning has documented in systems that learn from examples; and to reasoners who discover distinctions or possibilities they had overlooked.