Research Statement

Milan Mossé

My areas of specialization are moral and political philosophy. My work in these areas includes a series of papers on normative powers and reciprocity growing out of my dissertation, as well as two research projects on the ethics of AI, one about structural injustice and another about causal models of discrimination. I am a generalist at heart, and my work in value theory intersects with philosophy of race, philosophy of science, and social choice theory. These projects led me to develop two related research programs. The first lies in logic and concerns the difficulty of probabilistic and causal reasoning. The second lies in the philosophy of language and action and provides a unified model of means-end relations, which furnishes semantics for a wide range of normative and practical language, including deontic modals and intention ascriptions. Drafts of all papers under review (and most other papers) are available upon request.

1. Reasons by Request

Requesters seem to have what is, on reflection, a remarkable power: to create reasons at will. My dissertation explores this central but overlooked dimension of our normative lives. I explain how requests create reasons and how the values served by the power to request constrain its permissible use. In characterizing the valuable relationships sustained by requests, my dissertation develops a view of reciprocity that sheds new light on a wide range of further issues. These tasks fall into three papers, described in more detail in my dissertation summary:

- **How Requests Create Reasons** (under review) argues that requests create reasons by improving compliance with imperfect duties.
- **Why not Ask?** (under review) shows that requests can wrong others by creating reasons, which set back the requestee's interests.
- **Two Kinds of Reciprocity** (under review) introduces a distinction between material and relational reciprocity and applies it to a range of ethical and political issues.

In showing that my view explains the unfairness that can result from requests, I draw on my paper with Léa Bourguignon (**How to Count Sore Throats**, *Analysis*, 2025), which explains when and why fairness permits ties among competing interests to be broken. Two in-progress papers build on this work:

- My dissertation develops the idea that requesting is a normative power—an ability to create reasons at will, possession of which is explained by its value. A general challenge for the theory of normative powers is to make sense of the idea that the possession of an ability can be explained by its value. After all, it does not follow from the fact that it would be nice if I were able to cure cancer that I actually have that ability. In **Justified Wishful Thinking?**, I address the question of when exactly we can infer p from the fact that it would be valuable if p were true.

- My dissertation develops a theory of reciprocity with some surprising and substantial upshots for political philosophy. I explore some of these in my paper **Against Deportation**, where I argue that host countries have duties not to deport immigrants who reside and work illegally, including duties of reciprocity: states cannot permissibly accept the benefits of immigrants' compliance with the most costly duties that apply to citizens, while withholding basic benefits like residence associated with citizenship.

2. Artificial Intelligence and Structural Injustice

My research on the ethics of AI, which spans explainable AI, algorithmic fairness, and AI alignment, concerns *structural injustice*: I am interested in understanding how several decisions, which appear innocent in isolation, can create a system that is unjust. Three of my published papers fall under this topic:

- My work in explainable AI is concerned with how, even if (say) race and gender do not influence a decision when taken in isolation, an algorithm can still discriminate on the basis of race or gender, taken together with other factors. With Harry Sha and Li-Yang Tan, I developed a method for producing explanations of automated decision making (**A Generalization of the Satisfiability Coding Lemma and Its Applications**, *International Conference on Theory and Applications of Satisfiability Testing*, 2022), which resolves a problem posed by Quine over 70 years ago, and won the conference's Best Theory Paper Award.
- In a paper published in the premier algorithmic fairness venue (**Multiplicative Metric Fairness Under Composition**, *Foundations of Responsible Computing*, 2023), I prove that the most popular formal definition of equality of opportunity for individuals fails to protect against structural injustice, whereas a definition that has been largely overlooked does so. I show how definitions of "similar treatment for similar individuals" determine whether *groups* of individuals are treated unfairly, and whether several decision-makers that are fair in isolation from each other can *compose into a system* which is unfair.
- Leading methods for alignment like Reinforcement Learning from Human Feedback train AI to align with individuals' preferences, but doing so implicitly aggregates preferences in ways that may violate basic principles of fairness in social choice. With several co-authors, I proposed a framework for using social choice to align machine learning models' choices with human values (**Social Choice for AI Alignment**, *International Conference on Machine Learning*, 2024), which was the third most-cited paper in social choice in 2024. My interest in social choice extends beyond AI: in a paper with Wesley Holliday, Chase Norman, Eric Pacuit, and Cynthia Wang (**Stable Voting and the Splitting of Cycles**, under review), we resolve a conjecture about the relationship between two major voting rules.

I also published two papers on artificial intelligence in leading computer science venues, while I was completing my M.S. in Computer Science at Stanford:

- **Conditional Negative Sampling for Contrastive Learning of Visual Representations** (*International Conference on Machine Learning*, 2021) explains why certain transformations on datasets (e.g. adding random noise) improve learning, while others (e.g. rotating an image) do not.

- **Zero Shot Learning for Code Education** (*Association for the Advancement of Artificial Intelligence*, 2019) provides a method for "zero-shot learning," i.e. learning without access to any examples. We use a fixed probabilistic grammar to generate samples at training time, and the learner uses this simulated dataset instead of a real one. This paper won the conference's Outstanding Student Paper Award.

3. Causal Models for Social Constructions

Direct discrimination occurs when a person's protected attributes, like race or gender, *cause* their worse treatment. This causal notion of discrimination is invoked by algorithmic fairness criteria, U.S. law, and correspondence studies in the social sciences. But it is difficult to spell out in a precise and plausible way, because basic features of social kinds like race and gender shine an uncomfortable light on the fundamental assumptions of causal inference. For example, a cause must be *modular*, i.e. sufficiently separable from other attributes to isolate its causal role. But gender is embedded in a nexus of social factors that resist isolated treatment: if gender is socially constructed, in what sense can it cause worse treatment?

Faced with such problems, some give up on the view that attributes like gender are social kinds, rather than biological ones. Others suggest that we give up on traditional causal models for discrimination. With Frederick Eberhardt, Thomas Icard, and Kara Schechtman, I developed a framework for modeling discrimination that gives us the best of both worlds: we can formally model claims about discrimination, making them empirically testable, but *also* explicitly model how attributes like race and gender are socially constructed. The key is to view race and gender as *high-level abstractions* of their many constituents. In **Modeling Discrimination with Causal Abstraction**, we introduce this framework, which is now being employed by social scientists to test for discrimination over real-world datasets. In **Causal Models for Social Constructions**, we address a wider range of worries that arise when modeling social constructions. For example, social kinds are explained by social relations between people, and by self-reinforcing patterns of causation over time. But causal models standardly represent causes as attributes of individuals, rather than relations between them, and self-reinforcing causal relations seem at odds with the Causal Markov condition and the requirement of acyclicity. None of these features has a straightforward representation in standard causal modeling frameworks, but we illustrate with concrete examples how causal models of social kinds can be achieved.

4. How Hard Are Probabilistic and Causal Reasoning?

Causal reasoning seems indisputably harder than purely "correlational" or probabilistic reasoning: it requires that we perform deliberate experimentation and accept strong assumptions to justify causal conclusions. Even so, in a precise sense, causal and probabilistic reasoning are equally difficult. Duligur Ibeling, Thomas Icard, and I propose that many of the most prominent (i.e. frequentist and Bayesian) approaches to statistical inference can be understood, at least in part, as attempts to turn an *inductive* problem into a *deductive* one. We show how to computationally reduce problems about causation to equivalent ones about correlation: the computational complexity of deductive probabilistic and causal reasoning is the same (**Is Causal Reasoning Harder than Probabilistic Reasoning?**, *Review of Symbolic Logic*, 2022).

Several papers responding to our work extend our formal systems to include a special summation operator, a feature of Pearl's do-calculus. We show that suitably restricted, the operator does not change our results, but that when the operator is interpreted in the standard way, the do-calculus becomes undecidable: there is no procedure for causal reasoning, as it is traditionally modeled (**On probabilistic and causal reasoning with summation operators**, *Journal of Logic and Computation*, 2024). Positively, our work circumscribes the complexity of popular causal inference tools and shows that causal reasoning adds no computational cost. Negatively, we show that the standard way of formalizing the do-calculus makes probabilistic and causal reasoning impossible.

Logicians and psychologists often distinguish *qualitative* and *quantitative* probabilistic reasoning. But how exactly should this distinction be drawn? Duligur Ibeling, Thomas Icard, Krzysztof Mierzewski, and I use computational complexity to draw a robust distinction between qualitative and quantitative reasoning, which tracks independently important features of the corresponding logical systems, including finite axiomatizability (**Probing the Qualitative-Quantitative Divide**, *Annals of Pure and Applied Logic*, 2023). In a manuscript, **Reasoning about Confirmation**, which I presented at Carnegie Mellon's 2025 Probability Logic Workshop, I strengthen our prior results, and along the way resolve some open questions about the difficulty of reasoning about when a piece of evidence confirms a hypothesis.

5. Modeling Teleology

These are teleological explanations:

- We should compost to reduce methane emissions to slow global warming.
- John intends to go to the tailor to request thread to fix his jacket.
- The seedling needs to be watered to grow taller to reach the sunlight.
- The dishwasher runs its motor to splash suds to clean the dishes.

The situations described differ markedly, but they share a common structure. For example, we now know why John intends to go to the tailor, and how he intends to fix his jacket, just as we know why the dishwasher runs its motor and how it cleans the dishes. At the upcoming Eastern APA, Seth Yalcin and I will present our paper **Teleological Explanations**, which gives a unified semantics for teleological explanations. The heart of the paper is a new formal model for means-end relations. We show that this model furnishes novel semantics for a wide swath of evaluative and practical language, including deontic modals and intention and need ascriptions.