Research Statement

Adam Bouyamourn

I work in three broad areas. I am interested in *causal inference*, and especially *the application of machine learning tools to causal inference problems*; I am interested in *the social consequences and impacts of modern ML tools*; and I have a line of work in *selective inference and research incentives*.

Causal inference and ML

I published a solo-authored paper on LLM explainability at EMNLP ("Why LLMs Hallucinate"). The main result is a causal identification result. Semantic information and causal information both have a structural invariance property: if there is really a causal effect, changing the treatment changes the outcome; words have meanings, so changing a word changes the meaning of a sentence. Heuristically: LLMs hallucinate because they do not learn structurally invariant representations of the information they contain in their training data: they are stuck at the level of correlation, not (semantic) causation. Grounding solves the problem, because it imposes a kind of structural invariance on LLM outputs: can be output only if it is a paraphrase of a claim that has been verified.

My dissertation project applies modern optimization approaches to the problem of selecting sites in a multisite randomized control trial ("Where to Experiment?" – draft). I show that the Mean Squared Error of the estimate of the Population Average Treatment Effect is minimized when the mean vector of selected sites is as close as possible to the mean vector of population characteristics; and that the MSE of the estimate of the Conditional Average Treatment Effect is minimized when the distribution of sites is as close as possible to that of the population distribution of sites. I formulate these optimization problems as mixed integer programs, and implement them using commercial solvers. I contrast this to a recent approach that chooses a subset of sites to maximize the probability that the PATE is within the convex hull of sites. I show a "no free lunch" result applies to the problem of site selection: sites that are most representative at the average will generally not be representative of the full distribution of sites, and vice versa. In future work, I aim to apply distributionally robust optimization methods to this problem.

I published work on the experimental evaluation of reinforcement learning agents in a dynamic energy pricing problem with collaborators in computer science at ACM e-Energy ("Prospective Experiment"). This project involved assessing the performance of two RL agents at setting dynamic energy prices in order to induce office-workers to use less electricity at peak times. My role was to solve a causal inference problem: how can we compare the performance of two RL agents, given that they both required lengthy training periods? I proposed a within-subjects design, treating the RL agents as subjects.

I have work in progress with co-authors on covariate balance testing for natural experiments ("The Power of Prognosis"). My contributions to this paper were to provide necessary and sufficient conditions for balance testing to function as a falsification test for the assumption of as-if random assignment; development of the prognosis-weighted test statistic; and adapting the procedure to the case of assessing continuity of potential outcomes in a regression discontinuity design, as well as data analysis and package development.

During the PhD, I interned at Apple in the Health AI team, where I worked on a project studying the causal effect of exercise on atrial fibrillation risk. A puzzle in the cardiology literature is that exercise appears to both reduce and increase the risk of atrial fibrillation. By examining high-resolution exercise and heart monitoring data from the Apple Watch, I showed that, while exercise is broadly protective, large increases in exercise above baseline can cause atrial fibrillation events. This was an interesting case where the dose response curve had non-obvious structure: derivatives of the dose had effects with a different sign to that of the dose.

Social impacts of ML

With Alexander Tolbert (Emory) and Emily Diana (CMU), I have work in progress on algorithmic fairness, discrimination, and lending approval decisions ("Protected Characteristics and Lending Decisions"). Large historical gaps in loan approval rates exist between white and non-white applicants, even after accounting for the creditworthiness of the applicant. This leads to sorting, or "the Subprime Trap": non-white applicants are more likely to match with high-interest lenders, even when they have higher creditworthiness than white applicants. We use these stylized facts to motivate a model of algorithmic lending decisions in which race is used as an heuristic for lending decisions because other information about creditworthiness is not sufficient to explain credit risk. We show that a bank with risk management constraints can fail to lend even to customers with positive Net Present Value when their information about applicants is insufficiently precise. We derive optimal subsidies for banks to induce them to learn about the creditworthiness of minority applicants by making loans. We aim to submit this to the Symposium on the Foundations of Responsible Computing (non-archival), and eventually to FAccT.

We also have work in progress on estimating treatment effects when only a noisy proxy of treatment status is observed ("Latent Treatment"). Here we are interested in cases where the race of a user in a recommendation system is not known, but can be inferred, and hence used to evaluate the fairness of a given procedure.

A paper in progress, "Firm-Level Regulation of AI: An Information Perspective", considers a nested principal-agent problem. The state regulates a profit-maximizing firm that delegates its decision-making to an RL agent that chooses the distribution of output. The state has preferences over the distribution of output. The firm wishes to maximize profit, and chooses an algorithm to select the distribution of output. In this model, there is the possibility of agency loss between the state and the firm (the outer problem), *and* the firm and the algorithm (the inner problem). Notably, if the firm has perfect information about the behavior of the algorithm, the solution to the state's optimization problem is simply the solution to the principal-agent problem of regulating the firm to achieve an objective. The full problem is just the outer problem, and there is no 'special' problem of regulating AI. If the inner problem is nontrivial, however, the firm's uncertainty about the algorithm's decisions must be anticipated by the state in choosing an optimal contract for the firm. The moral is that AI-specific rulemaking is needed when the firm has imperfect information about the behavior of the algorithm, and highlights the value of auditing in safe AI deployment.

Selective inference and research incentives

I also use game theory and statistical theory to study research incentive problems.

A solo-authored paper, forthcoming at *Research and Politics*, "Collusive and Adversarial Replication", studies a situation in which social alignment between researchers and prospective replicators leads to a form of tacit collusion: Researchers can get away with conducting bad research, because social ties discourage prospective replicators from conducting ex post review of their research. While the paper studies the case of explicit misreporting by a Researcher, the general argument applies to 'softer' examples of bad practices. The moral is that adversarial ex post review is better for scientific progress.

A paper in progress, "Explore-Exploit: The Trade-Off Between Internal and External Validity", with Tak-Huen Chau, uses game theory to study what is meant by 'conflict' between internal and external validity. The basic idea is that a researcher choosing a study design faces an *explore-exploit* problem: they face a tension between optimizing a design for the probability of rejecting a given hypothesis; or they can maximize the generalizability or robustness of a given finding. This can be understood as a tension between representativeness and generalizability, and is referred to as 'the price of robustness' in the operations research literature. Methods that are optimal for minimizing average-case risk are not optimal for minimizing worst-case risk, and vice versa: there is no free lunch. We study pre-analysis plans to show that their adoption can incentivize researchers to opt for "safer" but less generalizable research designs.