These are loosely sorted from most- to least-compelling, with regard to how interesting/tractable I think they are, but I'm at least moderately excited about all of them.

Quantifying Incentives

We show that linear-world assumptions on Structural Causal Influence Models (SCIMs) [summary, recent paper, collaborators] allow bounding of value of information, response incentives, **unfairness**, value of control and **instrumental control incentive**. We demonstrate that under certain conditions Goodheart's law implies these bounds are realized; consequently, that quantitative estimates of fairness, etc. can be obtained given a linear-world SCIM.

Notably, these results are algorithm-agnostic, derived solely from the nature of the problem environment and the assumption that the algorithms are optimal. These methods are demonstrated on college admissions and recommender systems using real-world datasets.

Bounding of Causal Effects despite Structural Uncertainty

Recent work [Zhang, Tian, Bareinboim] outlines how to obtain "Partial Identification" of causal effects, effectively bounding the causal effects even when the effect itself is not identifiable. Here, we apply a similar technique, bounding causal effects where the source of uncertainty regards the structure of the causal model. Specifically, given a partially-directed acyclic graph (PDAG) we provide a general algorithm for bounding counterfactual quantities using canonical representations of SCMs, polynomial programing, and Gibbs sampling. Finally, we explore computational efficiencies obtained by considering only portions of the PDAG, reducing the curse of dimensionality.

Counterfactual Robustness to Distributional Shift

Recent neuroscience research suggests that humans may use counterfactual reasoning to adjust to distributional shift. We demonstrate conditions under which counterfactual knowledge is sufficient for recognizing and compensating for distributional shift. We weigh the merits of several estimation techniques, and highlight expected failure modes of these methods. These results are validated on both supervised and reinforcement learning settings involving varying amounts of distributional shift.

Cause-preserving Feature Extraction

Similar to how neural networks automate feature extraction for increasingly-large datasets, it should be possible to automate variable extraction for causal learning. But, the wrong choice of feature extraction may obfuscate all causal information. I postulate that there always exists a hierarchical partition of variables which maximally preserves causal information, and that we can get hints about this hierarchy from observational data. If so, there should exist approximate methods for reducing a dataset to a small set of cause-preserving variables (on which structural learning can then be applied).

Prior over Counterfactuals

As counterfactual reasoning becomes more widespread in machine learning, it becomes increasingly important to identify which priors over counterfactuals are reasonable to use. I postulate that there are “natural” priors on counterfactuals, from which traditional assumptions of minimality in structural learning can follow as a conclusion rather than an assumption. An added benefit is that a prior on counterfactuals could provide a more informative ordering on causal diagrams in an equivalence class, enabling better causal estimates despite structural uncertainty. To distinguish between choices of counterfactual priors, I seek testable implications which could be compared to real-world data.