Research Statement—Annie Liang

My work is at the intersection of economic theory, statistics, and machine learning. The papers discussed in this research statement are organized into three branches: (1) dynamic acquisition of correlated information, (2) the relationship between machine learning and economic modeling, and (3) the welfare implications of algorithmic predictions. My work contributes theoretical results as well as new analysis of data, and uses techniques and insights from economics and computer science. The methods developed in the papers in the second branch have applications especially in behavioral and experimental economics.

1. Dynamic Acquisition of Correlated Information

In a classic problem of sequential information acquisition, a Bayesian decision-maker (DM) allocates limited resources across different informational sources to inform an action that will be taken at an endogenously chosen time. His payoff depends on the action taken as well as on an unknown payoff-relevant state. Solutions for this problem exist in certain classes of models — e.g., the classic solutions of Wald (1945), Gittins and Jones (1979), and Weitzman (1979) — but exact results about information acquisition from *correlated* informational sources have proved elusive.[1]

And yet sources of information often are correlated. For example, if the payoff-relevant state is the COVID incidence rate in a given city and the decision-maker has access to measurements of COVID within different neighborhoods of this city, then we would expect incidence rates to be more similar in geographically adjacent neighborhoods. Or, if the payoff-relevant state is the fiscal cost of a new policy, and news outlets report on this cost with different and unknown biases, then we would expect news outlets to share similar biases depending on whether they are liberal or conservative.

In a set of papers, my coauthors and I show that dynamic information acquisition from correlated informational sources can be tractably analyzed in a canonical informational environment (Gaussian information). To understand these results, first observe that one possible strategy for information acquisition (neglecting dynamic considerations) is to acquire information each period from the source that maximally reduces uncertainty about the payoff-relevant state. We refer to this as the *myopic* rule, as it corresponds to a decision-maker who mistakenly believes each period to be the last possible period of information acquisition. The main inefficiency of myopic planning is that it neglects potential complementarities across signals. A signal that is uninformative on its own can be very informative when paired with other signals; thus, repeated (myopic) acquisition of the best single signal need not result in the best sequence of signals in general.

But in **Liang, Mu, and Syrgkanis** (*Proceedings of the 2018 ACM Conference on Economics and Computation*, 2018), we show that the myopic rule eventually coincides with the optimal rule at all sufficiently late periods. And in **Liang, Mu, and Syrgkanis** (*Econometrica*, 2022) we sharpen this insight when the decision-maker acquires information in continuous time. We show that under a weak condition on the decision-maker's prior belief, the optimal dynamic information strategy is precisely the myopic strategy from the beginning. This optimal information acquisition strategy

[1] An exception is a model introduced by Callander (2011), where the available signals are the realizations of a single Brownian motion path at different points.

is moreover history-independent (i.e., the entire path of information aqcuisition is known from the beginning) and robust across decision problems. As a consequence of these properties, the optimal stopping time and information acquisition strategy do not have to be solved for jointly in this problem; the optimal information acquisition strategy is the same no matter when the decision-maker plans to stop. We apply our characterizations to derive new substantive results in settings motivated by particular economic questions—for example, extending a result from Fudenberg et al. (2018) relating the speed and accuracy of decisions in binary choice problems, and solving for the equilibrium in a game between competing news sources.

Why is the myopic rule optimal in this environment? The key argument is that under our given conditions on the prior belief, intertemporal tradeoffs do not arise: the best way to acquire information for a decision tomorrow is also the best way to acquire information for a decision a week from now, and so forth. Formally, we show that there exists a history-independent dynamic strategy, which minimizes the decision-maker's posterior uncertainty about the payoff-relevant state at every moment of time. This argument applies and generalizes one of the few results that provide a Blackwell ordering of sequential information acquisition strategies (Greenshtein, 1996).

The previous two papers impose an important restriction on the informational environment, which is that the payoff-relevant state can only be fully learned if the decision-maker eventually acquires information from every source. In **Liang and Mu** (*Quarterly Journal of Economics*, 2020) we relax this assumption to permit redundancies across sources, so that a strict subset of sources can be sufficient for long-run learning of the payoff-relevant state. Within this more general environment, we consider information acquisition by a sequence of short-lived decision-makers who each choose one source of information, e.g., researchers who sequentially choose among different experiments for learning about a scientific parameter of interest. Different from the classic social learning model (Banerjee, 1992; Bikhchandani et al., 1992), we suppose that agents pass down their information across periods, thus turning off the inference problem that is essential to incomplete learning in the information cascades literature.

We build on Borgers et al. (2013) to define the notion of a *strongly complementary set* of information sources, and show that every strongly complementary set of sources is a long-run absorbing set (starting from some set of priors). Intuitively, because of the internal complementarities across sources within a strongly complementary set, acquisition of information from those sources reinforces the incentives to return to that set. When there is a unique strongly complementary set (as is the case in the information environments studied in Liang et al. (2018) and Liang et al. (2022)), then efficiency is obtained: agents eventually discover the best sources of information, and sample from these optimally. But when there are multiple such sets, then agents may fall into *learning traps* and acquire information inefficiently slowly, with welfare losses for society that can be arbitrarily large. We conclude by suggesting interventions that can prevent decision-makers from falling into these traps; for example, policymakers can restore efficient information aggregation by reshaping agent payoffs so that they are rewarded for information acquired over multiple periods. This observation is consistent with practices that have arisen in academic research, including the evaluation of researchers based on advancements developed across several papers.

2. The Relationship between Machine Learning and Economic Modeling

Machine learning algorithms have demonstrated striking success in prediction, including of certain economic variables, but are hard to interpret and thus tend to explain very little about underlying forces. Economic models, in contrast, often achieve goals such as unifying behavior across economic domains by identifying general abstractions, conveying narratives that shape our understanding of those behaviors, and providing frameworks for making predictions in new environments. In a sequence of papers, I study the question of whether black box methods can nevertheless be useful to an economic modeler whose goals extend beyond prediction. These papers ask: What properties structurally distinguish economic models from black box algorithms? When should we prefer economic models over black boxes for prediction, or the reverse? What are opportunities for using the two to build on one another—for example, can black box methods help us to build better intepretable models and predictions of behavior?

In **Fudenberg, Kleinberg, Liang, and Mullainathan** (*Journal of Political Economy*, 2022), we propose using machine learning (ML) algorithms to benchmark the predictive performance of economic models. The basic challenge here is understanding what constitutes a "good" predictive performance. For example, is an R^2 of 0.6 a success for the model? The raw prediction error is difficult to interpret, because it confounds error from two distinct sources: A model could predict poorly because (1) it has failed to capture important regularities relating the observables to the outcome; or (2) there is intrinsic randomness in the outcome conditional on the observables.

For a concrete example of the difference, consider a model that predicts information diffusion on a network based on a structural model of how individuals communicate on that network. If this model predicts poorly, one possibility is that the structural model is not leveraging the correct statistics of the network—in this case, another model using different statistics of the network (e.g., a different centrality measure) would lead to better predictions. The other possibility is that network structure alone is not sufficient to predict information diffusion, since there is substantial irreducible noise in that outcome conditioning on the network structure alone. In this case, substantial improvements in prediction would require the new model to be built on other features as well (e.g., demographic data about individuals in the network). Our proposed measure of *completeness* compares the predictive accuracy of the model to the best achievable accuracy given the observables. We show that the best achievable accuracy can be accurately estimated for a diverse range of laboratory data sets, using a simple "lookup table" algorithm that nonparametrically searches the space of possible models.

For problems in which our existing models are not yet complete, can machine learning techniques help us to identify new regularities in behavior, and/or guide us towards practical extensions of existing models? In **Fudenberg and Liang** (*American Economic Review*, 2019), we investigate these questions in the context of predicting how subjects play the first time they encounter a new normal-form game. We consider an aggregated dataset of initial play from six laboratory experiments, and show that a machine learning algorithm indeed predicts better out-of-sample than standard economic models when evaluated on these games. Of particular interest are the games on which play is well-predicted by the machine learning algorithm but poorly predicted by the economic models. Play in these games turn out to have a common regularity, which was not captured by the economic models. We formulate a one-parameter extension of the most predictive economic model

(Level-1) to include this regularity, and find that our new model (which we call Level-1(α)) predicts as well as the black box algorithms.

The strong performance of the new model is interesting in its own right, but leaves open the question of how robust its performance is beyond this specific set of experimental games. To study this, we need data on play in new games. We next propose using machine learning to guide experimental design by directing us towards new games on which the Level-1(α) model may fail. Specifically, we first train an algorithm to predict how well the Level-1(α) model will perform on an arbitrary game; then, we generate a large data set of games on which the algorithm predicts that Level-1(α) will not perform well. We have subjects on the platform Mechanical Turk play these "algorithmically-generated" games, and find that Level-1(α) indeed predicts behavior poorly in these games. In contrast, the predictive accuracy of machine learning algorithms is as high as on the original games. This suggests that behavior in the new games is not fundamentally unpredictable, but rather follows regularities that are not captured by the Level-1(α) model. Studying a highly constrained machine learning algorithm (specifically, a 2-split decision tree), we discover a new strategic property that is a good predictor of which action will be played in these algorithmically-generated games.

The previous papers build towards more complete models, with the understanding that a complete model is a desirable end goal. But one way in which a model might be very complete is by being very flexible (potentially not falsifiable!). Indeed, machine learning algorithms trained on large quantities of data achieve high completeness precisely for this reason. When economic models fit the data well, we need to understand whether this is because the model is precisely tailored to capture important structure in the behavior of interest, or if (similar to the black box algorithms) the economic model fits the data well simply because it is very flexible.

Fudenberg, Gao, and Liang (accepted at *Review of Economics and Statistics*, 2023) proposes a general computational method for quantifying the restrictiveness of an economic model, based on how well the model fits randomly generated data. Importantly, this measure can be computed without the guidance of analytical results about the empirical content of those models.[2] We apply the measure to several economic domains, and show that some economic models with a small number of parameters in fact impose very few restrictions on data. Their good fit to the data may thus be a consequence of their flexibility, rather than having placed the "right" restrictions. Combining our measure with the previous measure of completeness delivers a Pareto frontier, where models that rule out more regularities, yet capture the regularities that are present in real data, are preferred. We illustrate this Pareto frontier in an economic application and find that some economic models are Pareto-dominated—i.e., another model exists which is simultaneously more complete and also more restrictive.

If an economic model is not very restrictive, it runs a risk of overfitting to small datasets just the same as black box algorithms. A common approach in statistics and machine learning for preventing overfitting is to bias towards simpler models by penalizing the complexity of the learned model (suitably defined). **Liang** (*Journal of Economic Theory*, 2019) takes such an approach to problem of rationalizing choice data.

[2]This differentiates our measure from classic measures for the complexity of a function class, such as VC dimension, Rademacher complexity, or metric entropy (see e.g., Hastie et al. (2009)).

Empirical choice data typically contains inconsistencies, i.e., observations that cannot be rationalized as maximization of a single preference. A well-known economic model which is not falsifiable is rationalization of choice data using multiple preferences.[3] Specifically, the classic Kalai et al. (2002) approach interprets these inconsistencies as emerging from preference heterogeneity, and seeks the smallest number of preferences that perfectly rationalizes the data. Every data set can be rationalized in this way, but if some inconsistencies in the data are due to choice errors (rather than preference heterogeneity), then the Kalai et al. (2002) approach will overfit to the data, treating choice errors as evidence of additional preferences and leading to worse predictions out-of-sample. I propose instead modifying the Kalai et al. (2002) approach by minimizing a weighted average of the number of inferred preferences and the number of unexplained observations. That is, while it is always possible to rationalize a larger fraction of the data by adding an additional preference, if the additional gain in explanation is small we might still choose the model with fewer preferences. I show that when the choice data is generated from imperfect maximization of a set of preferences, and the underlying set of preferences is in fact small, then this approach exactly recovers the number of underlying preferences with high probability.

Finally, one reason to prefer restrictive models that fit the data well is because of an intuition that such a model has captured the "right" structure and is thus more likely to generalize across different economic domains (e.g., if we estimate parameters for a pricing model on one population of individuals, and use it to guide pricing in another). **Andrews, Fudenberg, Lei, Liang, and Wu** (working paper, 2023) provides an approach for assessing generalizability directly. Specifically, we derive finite-sample forecast intervals for a model's out-of-domain error, i.e., its performance when trained on data from one economic domain and tested on another. These forecast intervals are valid under an assumption that the distributions governing data in different economic domains are themselves drawn IID from some underlying meta-distribution. We do not require any assumptions on the model class itself; thus, the forecast intervals apply to economic models and machine learning algorithms alike, and can be used to compare their generalizability.

We apply our results to the problem of predicting certainty equivalents for lotteries, where we interpret "domains" to correspond to subject pools. We show that machine learning algorithms slightly outperform economic models out-of-sample when we restrict to a fixed subject pool. But economic models outpredict machine learning algorithms when the training data comes from one subject pool and the testing data comes from another. These findings suggest that the economic models indeed generalize better across domains.

3. Welfare Implications of Algorithmic Predictions.

In early machine learning applications (e.g., classification of handwritten digits), machine learning algorithms were optimized purely to maximize the accuracy of predictions. Now that these algorithms also guide consequential predictions about people—such as who is creditworthy or who needs a medical procedure—objectives beyond accuracy have become relevant for their design and regulation.

[3]See for example Ambrus and Rozen (2013).

One emerging concern is the possibility that algorithms are very accurate for individuals in one social group while highly error-prone for another. Several recent papers have documented *disparate impact* of this form (Arnold, Dobbie, and Hull, 2021; Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2021; Obermeyer, Powers, Vogeli, and Mullainathan, 2019). Ideally, an algorithm would both have low disparate impact and also be accurate; in practice, there may be intrinsic tradeoffs between these goals. To examine this tradeoff, **Liang, Lu, and Mu** (working paper, 2023) defines and studies a *fairness-accuracy frontier*, which consists of those outcomes that are optimal for an objective function in a broad class reflecting different views on how to trade off fairness and accuracy. Our results identify a simple property of the inputs, group-balance, which qualitatively determines the shape of the frontier. Inputs are group-balanced if each group can be given the lowest error (among all groups) by using some algorithm built on these inputs.

We further study an information-design problem where the designer flexibly regulates the inputs (e.g., by coarsening an input or banning its use) to achieve certain fairness goals, but the algorithm that uses these inputs is chosen by an agent who exclusively values accuracy. In general, because of the misaligned preferences between the designer and the agent setting an algorithm, it could be optimal for the designer to completely ban a given input from use in the algorithm. But we show that when inputs satisfy our group-balance property, then banning group identity is strictly suboptimal for all designers regardless of their fairness preferences.[4] When group identity is an input, then an even stronger conclusion holds: It is strictly suboptimal for any designer to completely ban any informative input, since the designer can always do better by permitting use of some garbling of group identity together with this input. When applied, for example, to the context of college admissions, this result suggests that when group identity is a permissible input in college admission decisions, then excluding test scores is welfare-reducing for all designers with the power to flexibly regulate inputs. On the other hand, when group identity is not permitted as an input in college admissions decisions (as is now the case in the United States following the Supreme Court decision of *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College*), the optimal regulation of covariates (under some fairness preferences) may indeed involve completely excluding test scores.

In **Liang and Madsen** (forthcoming at *Theoretical Economics*, 2023), we study the disparate impact of algorithms through another lens: how the use of new data for forecasting affects economic incentives in settings with moral hazard. Our model builds on the classic "career concerns" framework of Holmström (1999), in which an agent exerts effort to improve an outcome used by a market to forecast his type. Differently from Holmström (1999), we suppose that the market additionally bases its forecast on auxiliary data consisting of covariates describing the agent, which are observed prior to his choice of effort. We show that incentives for worker effort shift as a result of this data collection, resulting in a change in the total amount of effort exerted by workers across the population.

The average direction of this change is determined by the persistence of the data's predictive power. In particular, forecasting from data on enduring worker attributes (for instance, demographic covariates) leads to a decrease in total effort across the population. Conversely, forecasting from

[4]This result is relevant to an ongoing policy debate regarding whether to ban group identities in certain algorithmic predictions, such as the use of race in medical predictions (Manski, 2022).

data reflecting short-run circumstances (for instance, indicators of recent financial shocks or illness) boosts total effort. But this average effect masks a redistributional implication, with new data potentially leading to excessive incentives for some agents and insufficient incentives for others, thus increasing variation in effort across workers. We show that (holding fixed a given change in average effort), inequality of this sort decreases aggregate social welfare, even though our objective function does not explicitly penalize inequality. The effect is strong enough that data which moves average effort toward the socially optimal level but generates sufficient inequality can reduce aggregate welfare.

Finally, **Iakovlev and Liang** (working paper, 2023) asks whether the use of big data to inform prediction about an agent's type is generally welfare-improving for the agent, or if there is something lost when the agent can no longer interact with a human being. We focus on a key contrast between human evaluators and black box algorithms: it is possible to provide context to a human through conversation, but not to a black box algorithm. For example, if a job interviewer asks whether an applicant has ever been arrested, and if the job applicant has previously been arrested for participation in a peaceful environmental protest, then that applicant can explain the nature of his arrest in addition to answering the question. But if an organization employs an algorithm that takes as input whether the individual has been arrested, and not the reason for arrest, the candidate cannot ask for the algorithm to be re-trained with this additional information.

We define the *value of context* to be the extent to which the agent can improve the evaluator's perception of him when given the opportunity to (truthfully) supplement the evaluation with new covariates. The value of context depends on how the covariates are related to the agent's type. We show that when this relationship is unknown and satisfies a symmetry condition, then the value of context vanishes in expectation as the number of covariates grows large. That is, although the value of context can be large for specific realized distributions, it does not typically matter. This result suggests that in contexts where the agent prefers accurate evaluations, and has sufficient uncertainty about what is relevant to the prediction problem, the agent should prefer an algorithmic evaluator who observes a larger set of covariates over a human evaluator to whom the agent can provide context.

References

Ambrus, A. and K. Rozen (2013): "Rationalizing Choice with Multi-Self Models," *Economic Journal*.

Arnold, D., W. Dobbie, and P. Hull (2021): "Measuring Racial Discrimination in Algorithms," *AEA Papers and Proceedings*, 111, 49—54.

Banerjee, A. (1992): "A Simple Model of Herd Behavior," *Quaterly Journal of Economics*, 107, 797–817.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992): "A Theory of Fads, Fashion, Custom, and Cultural Change as Information Cascades," *Journal of Political Economy*, 100, 992–1026.

Borgers, T., A. Hernando-Veciana, and D. Krahmer (2013): "When Are Signals Complements Or Substitutes," *Journal of Economic Theory*, 148, 165–195.

Callander, S. (2011): "Searching and Learning by Trial and Error," *American Economic Review*, 101, 2277–2308.

Fudenberg, D., W. Gao, and A. Liang (2023): "How Flexible is that Functional Form? Measuring the Restrictiveness of Theories," Working Paper.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022): "Measuring the Completeness of Economic Models," *Journal of Political Economy*, 130, 956–990.

Fudenberg, D. and A. Liang (2019): "Predicting and Understanding Initial Play," *American Economic Review*, 109, 4112–4141.

Fudenberg, D., P. Strack, and T. Strzalecki (2018): "Speed, Accuracy, and the Optimal Timing of Choices," *American Economic Review*, 108, 3651–84.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2021): "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *Journal of Finance*.

Gittins, J. C. and D. M. Jones (1979): "A Dynamic Allocation Index for the Discounted Multiarmed Bandit Problem," *Biometrika*, 66, 561–565.

Greenshtein, E. (1996): "Comparison of Sequential Experiments," *The Annals of Statistics*, 24, 436–448.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning*, Springer.

Holmström, B. (1999): "Managerial Incentive Problems: A Dynamic Perspective," *The Review of Economic Studies*, 66, 169–182.

Iakovlev, A. and A. Liang (2023): "The Value of Context: Human versus Black Box Evaluators," Working Paper.

Kalai, G., A. Rubinstein, and R. Spiegler (2002): "Rationalizing Choice Functions by Multiple Rationales," *Econometrica*, 70, 2481–2488.

Liang, A. (2019): "Inference of preference heterogeneity from choice data," *Journal of Economic Theory*, 179, 275–311.

Liang, A. and E. Madsen (2023): "Data and Incentives," Forthcoming at Theoretical Economics.

Liang, A. and X. Mu (2020): "Complementary Information and Learning Traps," *Quarterly Journal of Economics*, 135, 389–448.

Liang, A., X. Mu, and V. Syrgkanis (2018): "Optimal and Myopic Information Acquisition," in *Proceedings of the 2018 ACM Conference on Economics and Computation*, New York, NY, USA: Association for Computing Machinery, EC '18, 45–46.

——— (2022): "Dynamically Aggregating Diverse Information," *Econometrica*, 90, 47–80.

Manski, C. F. (2022): "Patient-centered appraisal of race-free clinical risk assessment," *Health Economics*, 31, 2109–2114.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan (2019): "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 366, 447–453.

Wald, A. (1945): "Sequential Tests of Statistical Hypotheses," *The Annals of Mathematical Statistics*, 16, 117–186.

Weitzman, M. L. (1979): "Optimal Search for the Best Alternative," *Econometrica*, 47, 641–654.