Research Statement

July 2026

Kevin He

I am a microeconomic theorist who studies learning and information in social settings. My work examines how learning and equilibrium behavior interact in environments with multiple strategic agents: observed data shape agents' choices, while those choices determine what data are generated and observed. Economic agents rarely learn from simple random samples. Firms infer demand conditions from market prices, which reflect competitors' strategic behavior. Consumers judge the quality of a new product from friends' choices in social networks and, increasingly, from algorithmically curated stories in social-media feeds. Researchers access sensitive information through data releases by statistical agencies that seek to protect privacy while promoting credible research. Information is therefore mediated by markets, networks, platforms, and institutions, each populated by actors with their own incentives. My research develops models of learning in such social and strategic environments. I ask when people form accurate beliefs and make well-informed decisions, when instead they make persistent or systematic mistakes, and how policy and institutional design can reduce those errors and improve welfare. Across several projects, I complement the theoretical analysis with randomized online laboratory experiments that test the models' behavioral mechanisms. I begin by outlining several recurring themes that connect the strands of my work.

The first theme is the feedback loop among beliefs, behavior, and data. Today's beliefs shape today's behavior, which generates the data that shape tomorrow's beliefs. A manager's records contain information only about applicants whom past managers chose to interview. A story's prominence in a social-media feed reflects how many earlier users shared it. Such feedback can entrench mistakes: managers who are overly pessimistic about the applicant pool may hire an early candidate and end their search too quickly, censoring information about potential future applicants and making their successors still more pessimistic. Social-media users surrounded by misleading viral stories may rationally believe them and continue to spread them. Much of my research asks when this loop is self-correcting and when it is self-reinforcing. In some settings, the answer turns on behavioral assumptions. Managers who commit the common statistical fallacy of expecting excessive streak reversals (the "gambler's fallacy") misinfer from predecessors' truncated histories, amplifying pessimism. Rational managers, by contrast, eventually infer the applicant pool's true quality and correct their predecessors' pessimism. In other settings, the answer turns on the structure of the learning environment. A social-media platform with a highly viral news-feed algorithm can generate self-perpetuating cycles of misinformation, whereas a less viral algorithm ultimately surfaces the truth even when early users receive misleading information.

The second theme of my research is that the welfare consequences of biases often depend on the learning environment. Suppose a group of people faces an unknown state, such as the quality of a new restaurant or the value of adopting a new technology. When individuals learn about the unknown state both from their private information and from their friends' beliefs in a social network, people who neglect correlation among others' beliefs can converge to a wrong consensus with positive probability, whereas rational learners always converge to the correct consensus. Changes in the network structure that have essentially no effect on rational learners can sharply reduce the accuracy of social learning for correlation neglecters. Yet a similar correlation-neglect bias can improve welfare in a different setting: duopolistic competition in which firms receive correlated signals about market demand. I show that a firm neglecting the correlation between its private signal and its rival's signal gains the commitment power to expand production more aggressively after favorable demand information, forcing its rational opponent to cede market share. But I also show that the bias becomes harmful when firms are uncertain about market price elasticity or when biased firms tend to compete with one another rather than with rational firms. Learners and their environments must therefore be evaluated jointly: the welfare consequences of both rationality and bias depend on the environment and the equilibrium that generates the data.

The third theme of my work is that learning failures, once understood as equilibrium phenomena, become design problems. Most of my positive results have normative counterparts: a designer can improve outcomes by changing the data-generating process while accounting for learners' incentives and equilibrium behavior. For information aggregation within organizations, I show that rewiring communication networks through mentorship programs or deliberate information silos can improve learning. For social-media platforms, I identify a content-neutral news-feed algorithm that preserves the benefits of social amplification while preventing epidemics of misinformation. For information institutions, I derive disclosure rules that maximize the usefulness of information subject to privacy constraints and show how dissemination noise can be repurposed to screen out self-interested data miners.

Much of my work lies at the interface of economics and neighboring fields such as computer science, statistics, and mathematics. Substantively, many of my papers study how digital technologies—including social-media platforms and generative AI—reshape the information landscape. Methodologically, I use tools from mathematics and statistics to build rich, tractable models in economic theory. For example, a comparative-statics result for the Gittins index with respect to a learner's utility function is instrumental in comparing how frequently different sender types in a repeated signaling game must experiment with different signals when they are uncertain about opponents' strategies. Stochastic-approximation techniques allow me to trace the evolution of content on a social-media platform. Results on sets of uniqueness from mathematical tomography characterize the most informative signals under a privacy constraint. I engage with these neighboring fields through both publication and presentation: eight of my papers have appeared, either in full or as extended abstracts, at the computer science conference Economics and Computation (EC), and one was selected for an oral presentation at the AI and machine-learning conference ICLR. Some papers also appear in interdisciplinary science outlets, including two in PNAS, and offer tools and conceptual frameworks of broader scientific interest.

Below I describe five areas of my research. The first combines theory and experiments to study social learning in networks and platforms. The second examines misspecified learning and behavioral information preferences, showing how statistical biases can persist and distort economic behavior. The third develops tools for information design under privacy constraints. The fourth provides learning-based foundations for equilibrium refinements in games. The fifth focuses on the emerging field of human–AI interaction.

1. Social Learning in Networks and Platforms

Related papers

1. Dasaratha and He, "Network Structure and Naive Sequential Learning," **Theoretical Economics**, 2020.
2. Dasaratha and He, "An Experiment on Network Density and Sequential Learning," **Games and Economic Behavior**, 2021.
3. Dasaratha and He, "Aggregative Efficiency of Bayesian Learning in Networks," accepted at the **Journal of Political Economy**; Best Paper Award, *ACM Economics and Computation (EC) 2021*.
4. Dasaratha and He, "Network Structure and Social Learning," **ACM SIGecom Exchanges**, 2021.
5. Dasaratha and He, "Learning from Viral Information," revise and resubmit at the **Review of Economic Studies**; presented at *ACM Economics and Computation (EC) 2024*.
6. Dasaratha and He, "Attention and Social Learning," Working Paper, 2026.

This first strand studies how people learn from one another in networks and platforms. Social learning is ubiquitous: consumers learn about product quality from reviews, voters about political candidates from news shared by peers, and workers about organizational norms from colleagues. In a long-running collaboration with Krishna Dasaratha, I study how well societies aggregate dispersed information and how the answer depends on structural factors—the observation network or platform algorithm—and behavioral factors—sophistication in inference and in understanding others' incentives. A common intuition is that more connections and observations should improve learning. My research identifies plausible settings in which this intuition fails. The value of observing another person's action depends on what that person saw, how they

processed it, and how the network correlates information across people. More observations can therefore harm learning by creating double counting, confounding, and self-reinforcing errors.

In "Network Structure and Naive Sequential Learning," we study sequential social learning about an unknown state by naive agents who treat each observed action as if it reflected only that neighbor's private information. In a Gaussian environment, this assumption yields a tractable linear representation of actions in terms of network paths. The formula lets us compute naive agents' exact probabilities of converging to the correct action in arbitrary weighted networks, extending a literature that largely classifies networks according to whether agents converge to the correct action with probability one. Mislearning becomes more likely as link density and integration across different communities increase, because denser networks create more opportunities to double-count the same signal. In networks with two partially segregated groups, divergent early signals can generate persistent disagreement. These results speak to concerns that denser online connectivity may reduce collective accuracy while entrenching polarization.

We test this mechanism in "An Experiment on Network Density and Sequential Learning." In a preregistered experiment, participants are assigned to groups of 40 and act sequentially. Each guesses a hidden binary state after receiving a private signal and observing the guesses of a subset of predecessors. In sparse networks, each participant has a 25% chance of observing each predecessor; in dense networks, the probability is 75%. The accuracy gains from social learning are more than twice as large in sparse networks. This comparative static is predicted by models of naive inference in which agents neglect correlation across observations, but is difficult to reconcile with rational-learning models.

In "Aggregative Efficiency of Bayesian Learning in Networks" (whose main results also appear in the invited survey article "Network Structure and Social Learning" for the computer science community), we return to rational learning and ask how efficiently networks aggregate dispersed information when agents are Bayesian. We identify a novel network-based obstruction: *information confounding*. Suppose an agent observes only two neighbors, both of whom observed an earlier common predecessor not seen by the agent. From the agent's perspective, this unobserved early action confounds the informational content of the neighbors' behavior: the network makes it impossible to incorporate both neighbors' private information fully without overweighting the early mover's private information. Rational agents solve a signal-extraction problem to draw the best possible inferences in the presence of confounding, but some information is inevitably lost. We introduce a tractable model that quantifies this loss and compares the efficiency of rational social learning across networks.

We study the canonical sequential social-learning model with a binary state under two richness assumptions. Agents receive Gaussian private signals about the state and choose from action spaces rich enough to reveal their beliefs. They observe their private signals and their network neighbors' actions; under our assumptions, their rational actions are log-linear functions of these observations. Each rational agent's action is distributed as though the agent had observed some (possibly non-integer) number of independent private signals. This *signal-counting interpretation* lets us measure any agent's learning outcome in units of private signals.

We first illustrate how poorly some network structures aggregate information. In the leading example, an agent has many neighbors who share common predecessors not observed by the agent. Observing *any* number of neighbors who share only one common predecessor is less informative than observing four independent signals. Thus, even limited confounding can produce arbitrarily large information loss. To assess the mechanism's empirical relevance, we calculate losses from rational sequential learning in Facebook friendship networks among students at 100 universities. The losses vary substantially across schools—from 25% to 74% of available signals—and much of this variation is explained by a standard network statistic.

Moving beyond finite examples, we define *aggregative efficiency* as the long-run fraction of private signals incorporated into social consensus in an infinite network. Our main results show that this efficiency can be arbitrarily small in several classes of *generational networks*. Agents are arranged in generations of

size K, and each agent in generation t observes a subset of predecessors from generation $t-1$ and possibly earlier. Suppose that every agent observes the same number of neighbors from the immediately preceding generation and no one from earlier generations, and that every pair of distinct agents in a generation shares the same number of common neighbors. For these networks, we derive an exact expression for aggregative efficiency. Regardless of generation size, social learning asymptotically accumulates no more than two signals per generation, so efficiency approaches zero as generation size grows. The expression also permits network comparisons: efficiency rises with the number of observations but falls with the number of common neighbors, quantifying the relevant trade-off. Even if agents observe all predecessors from *any* finite number of previous generations, the asymptotic number of signals aggregated per generation remains bounded above by four.

These results raise the question of whether any cohort-based network can sustain efficient learning. We consider two organizational changes that can improve outcomes. First, opening new channels of communication—for example, a mentorship program in which senior members share their experience with newcomers—can greatly improve organizational learning and nearly eliminate inefficiency. Second, creating information silos by partitioning some employees into insular groups that do not communicate with one another improves executives' information aggregation at the expense of workers' learning. An asymmetric network can improve learning in one part of an organization by degrading it elsewhere, a trade-off that may be desirable when some roles make especially consequential decisions. Unlike the behavioral results above, this finding shows that removing observations can improve outcomes even when all members are rational, because doing so alleviates Bayesian information confounding.

In "Learning from Viral Information," we study how a society learns from "viral" signals propagated through a population, with news shared on social-media platforms as the primary example. Users encounter stories through platform-generated feeds, in which prior users' engagements—retweets, upvotes, and similar actions—affect visibility. We ask how feed design, especially the weight placed on popularity, affects learning.

The paper's key methodological innovation is an equilibrium model in which agents interact and learn from one another's signals. Agents arrive sequentially and learn about a binary state. Each receives a conditionally independent binary signal about the state, representing a news story, and observes a sample of predecessors' signals, representing a news feed. The sampling rule interpolates between uniformly sampling past signals and sampling each signal with probability proportional to its *popularity score*, which rises with endorsements. A *virality weight* governs the emphasis on popularity: a higher weight makes popular signals more likely to appear in the sample. Agents are Bayesian and know the sampling rule, so they account for selection in the signals they observe. They then choose which sampled signals to endorse, preferring signals that match the true state given their beliefs.

A more viral sampling rule can improve information aggregation and strengthen beliefs because rational agents understand the selection process: a sampled signal is more likely to have been endorsed and is therefore more informative. Once the virality weight crosses a threshold, however, self-perpetuating misleading steady states emerge discontinuously where most circulating signals are wrong; users who mostly encounter those signals rationally come to believe them and continue endorsing them. Greater virality therefore creates a trade-off: society forms stronger beliefs but faces a risk of misinformation epidemics in which most people are confidently wrong. This is another setting in which more social observation and interaction—the number of signals each agent samples and endorses—can reduce welfare in a rational society by making a misleading steady state more likely.

These results have two implications for social-media platforms. First, the model predicts the distribution of popularity scores, corresponding empirically to the number of shares, likes, retweets, and similar engagements. For each steady state, popularity scores converge to a stationary distribution that we solve for explicitly. The distributions have power-law tails whose thickness is a simple function of model parameters and the realized steady state. Second, we ask how platform design can improve accuracy, a question with

potential regulatory implications. We identify a content-neutral modification: vary the virality weight over time, using a low weight in early users' feeds and a high weight in later users' feeds. We show that a simple equilibrium then achieves high accuracy without producing misleading steady states. Intuitively, the platform allows independent information to accumulate early in the discussion of a new issue and exploits the benefits of viral amplification later on.

Methodologically, the paper uses stochastic-approximation techniques to track the society's evolution from its initial condition, rather than the martingale-convergence arguments traditionally applied to the public-belief process in social-learning models. We apply these tools to an *equilibrium* model in which agents respond optimally to an evolving stochastic process. Similar techniques have been used in economics to study dynamics under behavioral heuristics—for example, by Benaïm and Weibull (2003) in evolutionary game theory and Arieli, Babichenko, and Mueller-Frank (2024) in naive social learning. Applying them when agents use equilibrium strategies is more difficult. The key is an equivalence: society converges to a misleading steady state with positive probability under equilibrium sharing strategies if and only if it does so when agents follow the simple strategy of sharing signals that match the majority of their observations.

"Attention and Social Learning," our most recent project, identifies a different behavioral friction. People often rely on information intermediaries, such as journalists, who have better access to primary sources but differ in their incentives for accuracy. Journalists at newspapers that strongly reward careful, accurate coverage are likely to devote more effort to verification than those at outlets oriented toward sensationalism. Do observers understand how such incentives affect the accuracy of information from different intermediaries? This understanding is crucial both for resolving disagreements among intermediaries and for deciding how much to fact-check claims from different outlets.

In a large preregistered online experiment with nearly 1,000 participants, some participants learn a binary state by performing an attention task under high or low accuracy incentives. Others predict their peers' accuracy from those incentives and aggregate answers from peers facing different incentives. Only a slight majority understands even the direction of the incentive effect, and only a minority places strictly greater weight on the better-incentivized peer when peers disagree. In an attention-substitution task, participants can use both their own attention and a peer's answer to infer the state. Participants fail to pay more attention when paired with a peer who faced weaker incentives, contrary to leading models of flexible, costly information acquisition. Even when others' attention incentives are transparent, most people do not consistently incorporate them into their beliefs and behavior.

2. Misspecified Learning and Behavioral Information Preferences

Related papers

7. He, "Mislearning from Censored Data: The Gambler's Fallacy and Other Correlational Mistakes in Optimal-Stopping Problems," **Theoretical Economics**, 2022.
8. He and Libgober, "Higher-Order Beliefs and (Mis)learning from Prices," **American Economic Journal: Microeconomics**, 2026.
9. He and Libgober, "Misspecified Learning and Evolutionary Stability," **Journal of Economic Theory**, 2025.
10. Duraj and He, "Dynamic Information Preference and Communication with Diminishing Sensitivity Over News," **Theoretical Economics**, 2024.

When people learn from data, they may depart from perfect Bayesian rationality because of behavioral biases, statistical fallacies, or psychological preferences over the timing of information. This second strand studies misspecified learning and behavioral information preferences. A unifying theme is that features of the learning environment that may be irrelevant to a rational Bayesian's beliefs or welfare can substantially affect a behavioral or misspecified learner. These features include parameters of a sequential-search problem, assortative matching in strategic interactions, and the lumpiness of information arrival. Accounting for such mistakes and nonclassical preferences therefore gives policymakers additional scope to improve welfare.

"Mislearning from Censored Data: The Gambler's Fallacy and Other Correlational Mistakes in Optimal-Stopping Problems" studies optimal-stopping problems, such as managers conducting sequential interviews or sellers fielding a stream of offers, when people expect systematic reversals in random sequences. Agents are uncertain about the underlying distribution—for example, the talent distribution in an applicant pool—and infer its parameters from data generated by predecessors in the same environment. Because agents stop when an early draw is good enough, observed histories are asymmetrically censored: they contain long streaks of below-average draws but no long streaks of above-average draws. A learner prone to the gambler's fallacy finds the negative streaks highly improbable and infers that the distribution is worse than it is. Society converges to excessively pessimistic beliefs about the mean, and agents stop too early. The distortion is self-reinforcing across generations: pessimistic agents lower their thresholds, censor data more severely, and pass an even more misleading sample to their successors.

This paper identifies a new mechanism for mislearning: the interaction of psychological bias and endogenous censoring in stopping problems. Both elements are essential. Agents without correlational mistakes learn the fundamentals correctly even from censored histories. Conversely, when there is no censoring and agents observe ex-post what would have been drawn in each period, even biased agents learn correctly. I also derive novel comparative statics for how the economic environment affects learning. Subsidies for longer search mitigate belief distortions among agents with the gambler's fallacy because they raise acceptance thresholds, generate less heavily censored histories, and thereby induce less pessimistic beliefs in successors. These comparative statics arise specifically when biased agents learn from endogenously censored histories: the subsidies neither affect rational agents' long-run beliefs nor help misspecified agents who face exogenous data.

Correlational errors reappear in "Higher-Order Beliefs and (Mis)learning from Prices," where we ask when they can be beneficial in games. We analyze an incomplete-information Cournot duopoly repeated under independent draws of the demand-curve intercept. Other parameters, including price elasticity, remain fixed across repetitions. In each period, firms receive private signals about the idiosyncratic demand shock but may misperceive the correlation between their own signal and their rival's signal—that is, they hold incorrect higher-order beliefs about demand. When firms dogmatically hold correct beliefs about price elasticity, neglecting signal correlation is beneficial. When they are initially uncertain about elasticity and infer it from realized prices, however, the correlation misperception distorts that inference and reverses the welfare conclusion: the bias becomes harmful in equilibrium. Thus, the same misspecification can have opposite welfare effects depending on what else agents learn from equilibrium data. The effects also depend on matching assortativity—whether misspecified entrants are matched uniformly with rational incumbents or assortatively with one another. Assortativity has no welfare effect without misspecification.

In "Misspecified Learning and Evolutionary Stability," we generalize beyond this bias and apply the indirect evolutionary approach to model selection. Agents with different *models* meet to play a stage game; a model specifies the feasible beliefs about game parameters and others' strategies. In equilibrium, each agent adopts the feasible belief that best fits the data and optimizes accordingly. We assess the stability of a resident model by comparing its equilibrium payoff with that of an entrant model. The analysis identifies conditions under which inferential flexibility is *necessary* for a misspecified model to invade a rational society: a correctly specified resident model can be destabilized by a misspecified entrant model that permits multiple feasible beliefs—and hence inference—but not by any model with a single feasible belief. Unlike agents with dogmatically wrong beliefs, entrants with misspecified models may perform well only when their population share is sufficiently large. Data generated in matches with residents can differ from data generated in matches among entrants, leading entrants to infer different beliefs and change their behavior. We apply these results to explain the emergence of analogy-based reasoning in centipede games.

In "Dynamic Information Preference and Communication with Diminishing Sensitivity Over News," we study how diminishing sensitivity to news shapes preferences over information and dynamic communication. A person's future consumption depends on an unknown state. In each period, the person receives information

about the state and experiences gain–loss utility from changes in beliefs about future consumption ("news utility"). Good news produces elation and bad news disappointment, but the emotional response is concave in the magnitude of the news. How does the person prefer to learn over time? And how should an informed agent communicate the state when seeking to maximize the person's expected welfare? Diminishing sensitivity is a classic but surprisingly understudied assumption in Kahneman and Tversky's (1979) prospect theory. We show that it generates novel, testable predictions about information preferences and communication.

The first prediction concerns gradual versus one-shot information. Existing models of behavioral information preferences typically predict that agents always prefer one form or the other. With diminishing sensitivity, by contrast, the same person can choose differently across situations depending on the ranking of consumption outcomes across states. For example, the model allows the same person to watch one political debate live (thereby choosing gradual information) but only read about the outcome of another debate in the newspaper (thereby choosing one-shot information), depending on how they feel about the candidates.

When an informed, benevolent sender communicates the state through cheap talk, the receiver's diminishing sensitivity creates a credibility problem. With sufficiently low loss aversion, the lack-of-commitment problem is so severe that every equilibrium is payoff-equivalent to babbling. Sufficiently high loss aversion can restore the credibility of good-news messages by increasing the future disappointment cost of lying in the bad state. Receivers with greater loss aversion may therefore earn higher equilibrium payoffs.

3. Privacy and Information

Related papers

11. He and Mu, "Differentially Private and Incentive Compatible Recommendation System for the Adoption of Network Goods," **ACM Economics and Computation (EC)**, 2014.
12. He, Sandomirskiy, and Tamuz, "Private Private Information," **Journal of Political Economy**, 2026.
13. Echenique and He, "Screening p-Hackers: Dissemination Noise as Bait," **Proceedings of the National Academy of Sciences**, 2024.

The third strand asks how institutions can protect privacy while preserving information's usefulness for economic decisions. The first two papers treat privacy as an exogenous constraint and ask how much information an informed party can disclose subject to it. The third shows that noise commonly added to data releases for privacy can also enhance research credibility by deterring data mining.

My interest in privacy dates to my earliest work. In "Differentially Private and Incentive Compatible Recommendation System for the Adoption of Network Goods," we study a social planner who makes product recommendations to a group of agents. Because the product generates network externalities, the planner can improve coordination by revealing agents' adoption statuses. We treat each agent's status as private and solve for the optimal recommendation mechanism subject to a privacy constraint based on differential privacy, a concept from computer science.

In "Private Private Information," we introduce *private private signals*. A signal is usually called private when only one agent observes it, but such signals may reveal information about one another through correlation, creating privacy concerns even without public disclosure. We define a private private signal structure as one in which the signals are independent of one another: each contains information about an unknown state of nature but not about the realizations of other signals. This requirement creates a trade-off: to achieve privacy, signal quality may have to be sacrificed. For instance, not all agents can have perfectly revealing signals about the state: otherwise, each agent could learn the state from their own signal and hence infer any other agent's signal realization through its correlation with the state. We characterize the informativeness of private private signals and identify structures that are *optimal* in the sense that it is impossible to make every signal more Blackwell informative without violating the privacy constraint.

For example, consider a recommender who seeks to write the most informative possible letter about a job applicant's productivity while revealing nothing about a sensitive attribute such as health status. If health and

productivity are independent, the recommender can reveal productivity fully. If they are correlated, a fully informative letter about productivity also reveals health information and violates the privacy constraint. The letter must therefore be independent of health status, even though both the letter and health may be informative about productivity. Together, a permissible letter and health status form a private private signal structure. Our results give a simple procedure for constructing the most informative permissible letter from the statistical relationship between health and productivity.

Methodologically, we establish a surprising connection to the field of mathematical tomography and the study of sets of uniqueness. A subset of $[0, 1]^n$ is called a *set of uniqueness* if it is uniquely determined by its one-dimensional marginals along coordinate directions. Understanding such sets has been an active area of research since the 1940s. In the case of a binary state, we show that all private private information structures with n signals can be identified with subsets of $[0, 1]^n$ and that the optimal ones (in terms of Blackwell informativeness) correspond exactly to sets of uniqueness. In the two-dimensional case—which corresponds to the case of two signals—the complete characterization of the sets of uniqueness is known (Lorentz, 1949) and leads to a simple description of the Blackwell–Pareto frontier. With three or more states, we establish an equivalence between optimality and a generalization of sets of uniqueness that we term *partitions of uniqueness*.

In the final paper in this area, "Screening *p*-Hackers: Dissemination Noise as Bait," we study how an institution can release data while improving the credibility of empirical research. Statistical agencies routinely inject noise into public datasets, usually to protect privacy. We identify another benefit: deliberate noise can screen out *p*-hackers—researchers who fit many models in search of statistical significance without ex-ante subject-matter knowledge. A researcher who mines the released data for publishable correlations is disproportionately attracted to patterns created by the noise, whereas hypothesis-driven research is largely unaffected. An agency that retains the clean data can therefore identify and filter out data-mined findings.

4. Learning in Games as a Foundation for Equilibrium Refinements

Related papers

14. Fudenberg and He, "Learning and Type Compatibility in Signaling Games," **Econometrica**, 2018.
15. Fudenberg and He, "Payoff Information and Learning in Signaling Games," **Games and Economic Behavior**, 2020.
16. Fudenberg and He, "Player-Compatible Learning and Player-Compatible Equilibrium," **Journal of Economic Theory**, 2021.
17. Clark, Fudenberg, and He, "Observability, Dominance, and Induction in Learning Models," **Journal of Economic Theory**, 2022.
18. Fudenberg, He, and Imhof, "Bayesian Posteriors for Arbitrarily Rare Events," **Proceedings of the National Academy of Sciences**, 2017.

Many of game theory's most widely used solution concepts embed assumptions about beliefs after "off-path" events—contingencies that should not arise if everyone follows equilibrium play. A firm may make an unexpected price cut, or a job applicant may present an unusual credential. How should others interpret and respond to such deviations? This fourth area starts from the classical premise that a game's equilibrium is the long-run outcome of a society learning from experience and asks which off-path beliefs can emerge through that process. The process of adjustment to equilibrium then disciplines equilibrium beliefs after off-path behavior and refines predictions about the game's outcome.

The learning microfoundation in these papers exemplifies the combination of endogenous data and multiple strategic agents that recurs throughout my research. The stage game is an extensive-form game with n player roles. Suppose that n large populations of patient, long-lived agents correspond to these roles. Initially, agents are uncertain about how typical opponents play. In each period, agents from different populations are randomly matched to play the stage game. After each match, they observe the realized path of play, which generally does not reveal opponents' complete extensive-form strategies. Forward-looking

agents maximize expected discounted utility, so young agents have an option-value incentive to experiment with different strategies to explore how opponents respond to different actions. We analyze the limit of steady-state learning outcomes as agents become more patient and long-lived. Actions that are "off-path" in equilibrium often occur with positive probability as rational experiments by inexperienced agents. The feedback agents receive after each match shapes experimentation incentives and can therefore affect equilibrium selection.

In "Learning and Type Compatibility in Signaling Games," we develop this approach for general signaling games, where equilibrium selection hinges on how receivers interpret off-path signals. Inexperienced senders who are uncertain about receivers' strategies experiment with different signals to learn how receivers respond. Because sender types have different utility functions, their incentives to experiment differ. We say one type is more *type-compatible* with a signal s than another if, whenever s is a weak best response for the less-compatible type against some receiver strategy, it is a strict best response for the more-compatible type against that strategy. This condition on the one-shot signaling game disciplines dynamic experimentation: in every steady state, the more-compatible type chooses the signal more often on average. We then show that, as agents become more patient and long-lived, receivers learn these relative experimentation frequencies and every steady state satisfies the *strong compatibility criterion*, which restricts beliefs about the relative likelihoods of more- and less-compatible senders after off-path signals.

Technically, the paper is tractable because each sender type faces a multi-armed bandit problem, with the stage-game signals as arms. We compare types' optimal dynamic behavior through their Gittins indices. We also draw on our paper "Bayesian Posteriors for Arbitrarily Rare Events," which identifies conditions under which the amount of data required to infer the relative likelihoods of two rare events is proportional to the inverse of the rare-event probability. In the learning-in-games model, the rare event is a receiver's observation of an off-path signal sent as an experiment by an inexperienced sender.

In "Payoff Information and Learning in Signaling Games," we extend this agenda by allowing learners to know their opponents' payoff functions, as refinement concepts typically assume. This knowledge can both restrict and expand the set of long-run learning outcomes because experimentation incentives depend on which opponent responses learners deem plausible.

"Player-Compatible Learning and Player-Compatible Equilibrium" extends the analysis beyond signaling games to refinements in more general games. Many refinement concepts invoke "trembles": players fail to best respond with small probability because they make mistakes. We propose Player-Compatible Equilibrium (PCE), which instead interprets trembles as deliberate experiments by agents learning how opponents play. PCE makes a unique and intuitive prediction in several examples where trembling-hand perfect equilibrium (Selten, 1975) and proper equilibrium (Myerson, 1978) leave multiple Nash equilibria. Moreover, unlike earlier tremble-based concepts, PCE is grounded in an explicit learning model. Both rational learning and the common heuristic of weighted fictitious play imply the same restrictions on trembles as PCE.

Equally important is understanding what learning cannot justify. Although learning implies forward-induction-like refinements in signaling games, that conclusion does not generally extend to other games. "Observability, Dominance, and Induction in Learning Models" shows that learning does not, in general, justify eliminating weakly dominated strategies. A rational learner may use a dominated strategy as a cheap "trial balloon" to discover whether a non-dominated but very risky action is a best response. We also study coarser feedback environments in which agents receive less information than they would from observing the realized path of play, showing how observability affects which equilibria learning can reach.

5. Economics of Human–AI Interaction

Related papers

19. He, Shorrer, and Xia, "Human Misperception of Generative-AI Alignment: A Laboratory Experiment," Working Paper, presented at *ACM Economics and Computation (EC)* 2025.

20. Arghal, He, Saeedi Bidokhti, and Sarkar, "Steering the Herd: A Framework for LLM-based Control of Social Learning," **International Conference on Learning Representations (ICLR) Oral**, 2026.

My most recent work on human–AI interaction revisits themes of misspecification and social learning.

In "Human Misperception of Generative-AI Alignment," we conduct an incentivized online laboratory experiment on how people perceive AI alignment in economic decisions. As AI increasingly serves as an advisor, agent, and intermediary, the central issue is not only whether it is accurate or aligned but whether people correctly assess its accuracy and alignment, and how those assessments shape trust and delegation. Participants make their own choices in problems involving risk, time preference, social preference, and strategic interaction, then predict choices made by AI on behalf of a human user. They overestimate alignment between AI and human choices: in every problem, average predictions about AI are closer to average human choices than to actual AI choices. Individual predictions are also strongly correlated with participants' own choices, indicating that they exaggerate alignment between their preferences and AI behavior. A stylized model shows that these misperceptions distort delegation and can generate unexpected welfare comparative statics, including lower welfare when AI becomes objectively more aligned with human preferences.

In "Steering the Herd," coauthored with student and faculty collaborators from Penn Engineering, we study algorithms that increasingly mediate information through targeted advertising, personalized search, and AI assistants. We introduce a model of controlled sequential social learning in which a planner chooses, at a cost, the precision of agents' private signals while agents learn from predecessors' decisions. We characterize the optimal precision policy and run simulations in which large language models act as both planner and agents. The LLM planner chooses a policy broadly consistent with the theoretical optimum while also displaying deviations consistent with non-Bayesian reasoning.

6. Conclusion

Economic agents rarely observe raw facts. They observe selected histories, prices, endorsements, recommendations, publications, and AI-generated summaries—outputs shaped by incentives and interpreted through models that may be incomplete or wrong. Understanding this process is central to economics because beliefs drive market behavior, strategic interaction, political decisions, scientific credibility, and technology adoption. I plan to continue developing tools to analyze this process and to design environments in which learning is more accurate, privacy is better protected, and information intermediaries are more trustworthy.

References

Arieli, Itai, Yakov Babichenko, and Manuel Mueller-Frank. 2024. "Sequential Naive Learning." Working paper.

Benaïm, Michel, and Jörgen W. Weibull. 2003. "Deterministic Approximation of Stochastic Evolution in Games." *Econometrica* 71 (3): 873–903.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47 (2): 263–291.

Lorentz, G. G. 1949. "A Problem of Plane Measure." *American Journal of Mathematics* 71 (2): 417–426.

Myerson, Roger B. 1978. "Refinements of the Nash Equilibrium Concept." *International Journal of Game Theory* 7 (2): 73–80.

Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4 (1): 25–55.