Designing Shared Information Displays for Agents of Varying Strategic Sophistication

DONGPING ZHANG, Northwestern University, USA
JASON HARTLINE, Northwestern University, USA
JESSICA HULLMAN, Northwestern University, USA

Data-driven predictions are often perceived as inaccurate in hindsight due to behavioral responses [51]. We consider the role of interface design choices on how individuals respond to predictions presented on a shared information display in a strategic setting. We introduce a novel staged experimental design to investigate the effects of interface design features, such as the visualization of prediction uncertainty and prediction error, within a repeated congestion game. In this game, participants assume the role of taxi drivers and use a shared information display to decide where to search for their next ride. Our experimental design endows agents with varying level-k depths of thinking [8], allowing some agents to possess greater sophistication in anticipating the decisions of others using the same information display. Through several large pre-registered experiments, we identify trade-offs between displays that are optimal for individual decisions and those that best serve the collective social welfare of the system. Additionally, we note that the influence of display characteristics varies based on an agent's strategic sophistication. We observe that design choices promoting individual-level decision-making can lead to suboptimal system outcomes, as manifested by a lower realization of potential social welfare. However, this decline in social welfare is offset by a slight reduction in distribution shift, narrowing the gap between predicted and realized system outcomes. This may enhance the perceived reliability and trustworthiness of the information display post hoc. Our findings pave the way for new research questions concerning the design of effective prediction interfaces in strategic environments.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; **Visualization design and evaluation methods**.

Additional Key Words and Phrases: behavioral game theory, congestion game, strategic decision-making, uncertainty visualization

ACM Reference Format:
Dongping Zhang, Jason Hartline, and Jessica Hullman. 2024. Designing Shared Information Displays for Agents of Varying Strategic Sophistication. In . ACM, New York, NY, USA, 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Technological affordances enable service providers to leverage historical data and offer users predictions from statistical models to assist decision-making [22, 40]. For example, ad marketplace owners present marketers with predicted outcomes in terms of clicks or ad placement by bid amount. Rideshare and taxi drivers, not to mention members of the public attempting to travel from point A to point B, consult predicted demand surges or traffic congestion from their App or Google Map to decide where to search for a ride or which route to take.

These everyday decision scenarios can be viewed as strategic settings of non-cooperative game theory, in which multiple *agents* use a shared information display provided by a *principal* to make decisions and lead to individual *payoffs* that depend on the agents' own choice and the choices of other agents.

Manuscript submitted to ACM

In principle, having access to predictions from information displays can benefit agents by providing them with exogenous, payoff-relevant information. However, in practice, the full benefit of information access may not always be realized. Consider a display predicting a taxi driver's chance of getting a pickup based on the number of other taxis on the road and historical data about where drivers tend to go. Such information displays could guide a driver's decision about where to head next. Yet, taxi drivers attempting to best respond to such a display might be taken aback when the displayed predictions are not realized. The problem lies in the fact that decision-making based on a shared information display becomes more challenging to use in multi-agent strategic settings: the system outcome formed by combining individual-level decisions is subject to *distribution shift* [26], where the predicted outcome shown on the display is inaccurate in hindsight because of agents' strategic responses to the displayed predictions.

A principal (i.e., service provider like a taxi company) faces limited options when confronted with persistent distribution shift resulting from behavioral reactions. Periodically retraining the model is often used in practice. Recent works in machine learning research [46, 51] propose exploring fixed-point solutions within the model retraining space to account for human behavioral responses upon viewing predictions. Drawing inspiration from information design [6], a game-theoretical approach involves selectively providing agents with payoff-relevant information to persuade behavioral change. However, the aforementioned solutions have drawbacks, as they can be costly (e.g., model retraining) or rely on stringent theoretical assumptions. We investigate an alternative approach through the lens of interface design and explore the design factors of shared information displays by examining how they can influence the individual-level decision-making and the aggregate-level system outcomes. Additionally, we explore the stability of these dynamics during repeated strategic decision-making over time.

We contribute the design and results of a large pre-registered online staged experiment using a repeated three-action congestion game based on the search-pickup dynamics of 2.1 million real-world taxi trips. In our experiment, we act as the principal, or the taxi company, while participants play as agents who are taxi drivers. As the principal, the taxi company's objective is to help drivers make good search decisions, ultimately leading to more pickups and improved overall efficiency. To accomplish this objective, the taxi company uses knowledge of how many drivers are on the road and historical data on supply and demand to train a statistical model and present the deduced flows and predicted pickup probabilities of different districts of a city through a visual interface accessible to all drivers.

We postulate that two design factors of a shared information display may be particularly influential in shaping agents' decision processes. First, we manipulate whether uncertainty in the predicted outcomes is visualized directly, as a salient depiction of uncertainty for predictions may encourage agents to fixate less on a single best response. Second, we manipulate whether *realized prediction error*—the difference between the predicted outcome and what actually happened—is visualized, as seeing *how* predictions tend to be wrong may help some agents make better use of an "inaccurate" information display. To study how these factors affect the strategic decision-making of agents with realistic variation in their levels of sophistication, we use a staged experimental design, in which we endow each agent with a level-k depth of thinking according to a Poisson Cognitive Hierarchy Model (Poisson-CH model) [8]. The Poisson-CH model posits that behavioral responses in a strategic setting can be explained by assuming a population comprised of agents who exhibit varying levels of sophistication in how they anticipate other agents' responses. For example, a level-0 agent behaves non-strategically; a level-1 agent attempts to best respond to a population of only level-0s; a level-2 agent attempts to best respond to a fixed mixture over level-0s and level-1s, etc.

Our results shed light on the interplay of design elements with strategic outcomes and underscore the challenges of designing shared information displays tailored for agents with varying levels of strategic sophistication. We discover that incorporating post hoc decision feedback through visualizing realized prediction error can help more strategically

sophisticated agents (i.e., level-2s) make more informed decisions. When decisions are combined to construct the system outcome according to the Poisson distribution used to define the frequency of all levels, design manipulations that can improve individual-level decisions can lead to decreasing social welfare relative to the system optimal over repeated decisions. Hence, desirable outcomes at the system level can be opposed to those at the individual level. At the same time, the decrease in social welfare is accompanied by a slight reduction in distribution shift, narrowing the gap between the predicted outcome and the realized system outcome, and hence improving perceived reliability and trustworthiness of the information display post hoc. We find that these results are robust across two close replications of our experiment, in which we varied the order of decision scenarios and the level distribution of strategic sophistication. We conclude by discussing how our work motivates new research questions around how communicating prediction uncertainty and error affects trust and reliance on shared information displays in strategic environments.

2 RELATED WORK

2.1 Information Design in Congestion Games

We study strategic decision-making in congestion games, which represent a broad class of non-cooperative games. Each action of the game represents a congestible good (e.g., local demand or traffic bandwidth) and is associated with a cost function, which incurs cost that increases with the number or fraction of agents who chose the same action [54, 55]. In our experiment, the principal's provision and manipulation of displayed information resembles the problem of information design in Economics, which studies how a principal can selectively provide payoff-relevant information to influence agents' behavior so as to better achieve the principal's objectives [6]. Previous work by Das et al. [15] shows that information design can mitigate congestion and improve social welfare in a congestion game. In contrast to selectively releasing information to influence decision-making, our work considers a scenario in which a principal is committed to providing all agents equal access to a shared information display but faces the choice of whether to present agents with prediction uncertainty or post hoc decision feedback on realized prediction error. Our work complements information design in economics by evaluating how the provision of prediction uncertainty and error impact both individual-level decision-making and aggregate-level system outcome in repeated decision-making.

2.2 Strategic Sophistication in Game Theory

Standard practice in game theory assumes that agents are fully rational and capable of error-free calculations using payoff-relevant information. However, behavioral economists view agents' utility maximization problem through the lens of bounded rationality [60], in which agents, typically constrained by limitation of both knowledge and computational capacity, tend to satisfice and adapt during decision-making [34, 58, 59]. This perspective has led to the development of several behavioral models, such as Cognitive Hierarchy Models [8], Prospect Theory [35], and Quantal Response Equilibrium [45], that aim to explain and model the underlying mechanism that dictates behavioral agents' decision-making. Our work adopts a Poisson-based Cognitive Hierarchy Model, which has been extensively studied by empirical game theorists (e.g., [17, 20, 62]), by endowing strategic sophistication to agents in a congestion game through a level-k framework. The Poisson-CH model defines the frequency distribution of agents' levels by a Poisson distribution. All agents within the level-k framework are considered to be myopic; they assume they are the most sophisticated agents in action and that all other competing agents are distributed according to a normalized Poisson for levels between 0 and $k-1$. We use the level-k framework to understand how interface design features can affect agents differently depending on their levels of strategic sophistication.

2.3 Information Displays for Strategic Decision-making

Research on uncertainty visualization tackles questions such as how to incentivize uncertainty communication [30], how to depict uncertainty information (e.g., [13, 19, 37]), and the challenges of evaluating uncertainty displays even for individual decision-making [31]. However, design goals for data and uncertainty visualization have traditionally prioritized individual outcomes, aiming to maximize perception or individual decision quality. Designing interfaces solely to achieve these objectives may not necessarily align with aggregate-level desiderata like greater efficiency or social welfare. Our work tackles novel questions related to how visualizing uncertainty can impact strategic decision-making.

Our work is closely related to Kayongo et al. [38] who propose the concept of visualization equilibrium. A visualization is in equilibrium if the system outcome observed from agents' decision-making mimics or closely approximates the distribution shown in the display. By studying initial play (i.e., decisions made with no feedback) in a two-action congestion game, they demonstrated how an outcome that a principle might desire for the system, such as a Nash Equilibrium, cannot be achieved by visualizing that outcome, but one can estimate an equilibrium by finding a displayed prediction that matches the realized outcome. Hence, their experiment differs from ours in two important ways: (1) the predictions that the principal provides in their set-up are not constrained by any exogenous information (they can be entirely fictitious), whereas we study a setting in which the principal's predictions are more realistically constrained by real-world taxi behaviors, and (2) they study initial play where agents do not observe any information about the realized outcome (either their choice or the aggregate system outcome) after using a display to make a decision.

Kayongo et al. [38] also proposes a hypothesis on the impact of visualizing prediction uncertainty on agents' ability to anticipate other agents' actions. They suggested that using a display that can make prediction uncertainty more salient, such as through animated hypothetical outcomes [32], may pose challenges for agents in predicting how others will react to the same display. While they provide weak evidence to support this hypothesis, our study aims to further investigate its validity through pre-registered experiments. From a qualitative analysis of participants' reported strategies, they find reports that suggest agents' decision-making behavior might be characterized by varying levels of strategic sophistication when it comes to anticipating other agents' responses, though this analysis is speculative, as they did not attempt to measure or endow strategic sophistication in their experiment. In contrast, we explicitly endow different levels of strategic belief according to a Poisson-CH model to understand design factors by level and vary the distribution over levels to check the robustness of our results.

3 ONLINE EXPERIMENTS

3.1 Overview

We conduct a large pre-registered[1] between-subjects repeated measures experiment on Prolific and two robustness checks in which we replicate the main experiment but change a single assumption. We study a three-action congestion game modeled after real-world situations that involve selecting between congestible goods [55], in which agents (i.e., participants) act as taxi drivers and are asked to (1) *anticipate* other participants' actions and (2) *decide* where to search for their next ride from three districts using a shared information display. Participants can use the display in this situation to help them maximize their payoffs—or chance of getting a pickup—by accounting for the displayed predictions in their decision processes. However, since all participants can access the same information display, whether or not a participant's decision can result in a pickup depends on the decisions of other participants. When more participants choose to search in the same district, the predicted chance of getting a pickup in that district is lower

[1]Pre-registration: https://aspredicted.org/pp8s8.pdf

on average. Because agents who use information displays in strategic settings such as congestion games are often long-lived, we observe learning from repeated plays over 15 trials.

The critical component of our congestion game is access to a *counterfactual model* that can compute realistic payoffs for players after decision-making. To inject realism into our setting and to evaluate decisions under exogenous predictions that are not fabricated, we analyzed the search-pickup dynamics of 2.1 million Chicago taxi trips and trained a counterfactual model that has a functional form of $pickups = f(flow)$. The model is designed to predict 9 AM taxi pickups (i.e., the designated prediction timestamp) in three Chicago Community Areas, corresponding to each action (i.e., district) of the game, given a discrete flow distribution of drivers going to search over the three districts. We use this model to (1) create predictions modeled after taxi search flows and (2) evaluate participants' decisions.

Decision-making in strategic settings such as the one we study has been found to be well described by assuming that players can vary in how sophisticated they are when anticipating other players' actions [8]. The shared information display we define embodies the assumption that drivers will act *non-strategically*, relying on their prior driving experiences *without access to the display*. We define these drivers as level-0s (L0s) according to the Poisson-CH model [8], and simulate the behaviors of these drivers in each decision scenario based on their past search preferences using the taxi data (see Appendix A.3).

We endow our study participants with either level-1 (L1) or level-2 (L2) depth of thinking according to a Poisson distribution that characterizes the level frequency over the participant population. Participants who are endowed with L1 belief assume all other participants are L0s, and therefore the displayed prediction closely resembles the *level-specific outcome* against which they will be scored. L2 participants who believe the population is composed of a fixed mixture of L0s and L1s are expected to make their decisions by combining the information about L0s from the information display with their beliefs about how L1s will attempt to best respond to that display. L2s are scored against a *level-specific outcome* that matches their beliefs, simulated by sampling L0s' decisions from the historical taxi data and L1s' decisions from L1 participants' responses according to L2's endowed belief that there is fixed mixture of L0s and L1s. Beyond allowing us to study interface effects over realistic variation in participants' levels of strategic sophistication, level endowment enables us to observe how the influence of interface design factors may depend on the extent to which the displayed prediction is aligned with the distribution of behavior that produces the participants' payoffs.

In our manuscript, we employ several specialized terms to delineate our setting and present our findings. For clarity and ease of reference, a glossary of these terms is provided in Table 7 of Appendix C.

3.1.1 Experimental Manipulations. Similar to Kayongo et al. [38], we vary whether the prediction uncertainty is visualized directly by showing participants either **static point estimates** or animated **hypothetical outcomes plots** (HOPs) [32]. HOPs are a frequency-based uncertainty visualization technique that presents a finite set of samples from a distribution, or predictions in our context, through a sequence of animated frames. Previous visualization studies suggest that HOPs can yield more accurate judgments than error bars [28, 32, 36] or other static methods such as static ensembles and violin plots [32]. We expect that visualizing static point estimates will lead to less variance in decisions from L1s and L2s, whereas visualizing uncertainty more saliently via HOPs will increase variance in decisions by helping participants recognize that the L0 decisions are not deterministic.

In each trial, participants are provided with post hoc feedback after decision-making. **Bandit** feedback only informs the participants if they received a pickup based on their decision. **Full** feedback informs the participants if they received a pickup based on their decision and visualizes the *realized prediction error*, or the difference between the predictions they saw prior to making their decisions and the *level-specific outcome* used to evaluate their decisions. Full feedback also

visualizes the participant's *anticipation error*, or the difference between the participant's anticipation of the *level-specific outcome* and that is realized. By varying the feedback structures, we are primarily interested in assessing whether visualizing realized prediction error can help participants anticipate the prediction error as a result of strategic behavior.

Table 1. Treatment Conditions

Interface	Uncertainty Display	Feedback Structure
1	Static	Bandit
2	Static	Full
3	NetHOPs	Bandit
4	NetHOPs	Full

We vary the uncertainty display and the feedback structure between subjects, resulting in four conditions outlined in Table 1. For each condition, we evaluate participants's performance using two individual-level responses: a binary indicator of whether the participant *best responded* to the display assuming their level-specific beliefs are correct, and a quantitative measure of their *anticipation error* in anticipating other participants' decisions. Additionally, we also simulate the more realistic *system outcome* for each decision scenario, calculated by assuming that the population consists of a fixed mixture of L0s, L1s, and L2s, governed by the Poisson distribution used to define the frequency of all levels. We calculate the *achieved social welfare*, representing the fraction of the total possible social welfare for that decision scenario (i.e., trial) that was achieved, and the *distribution shift*, representing the difference between the displayed prediction and the system outcome. Figure 1 demonstrates the features of our experiment design, describing how a distribution over all participants' levels of sophistication gives rise to a data collection arm where participants' *level-specific feedback* is generated given the endowed level-k beliefs and how we combine decisions to aggregate system outcomes. We describe each step of our experiment with references to Appendix and supplementary material.

3.2 Methods

3.2.1 Endowing Levels of Strategic Sophistication. A core aspect of our experimental design is the integration of the Poisson-CH model. In our main experiment, we use a Poisson(λ) where $\lambda = 1.5$ to define the frequency of levels in decision scenarios (i.e., trials) consisting of N participants, based on the findings of Camerer et al. [8] who analyzed the interplay of level distributions and results of nearly 100 games. Prior work [8, 11, 63] suggests that human players tend to perform one to two depths of thinking in strategic games, so we truncate the Poisson(1.5) and re-standardize its level distribution to a maximum level-k where $k < 3$. Truncating levels in the context of a Poisson-CH model is a common practice within the level-k framework (e.g., [53, 65, 66]). The rationale for excluding levels beyond $k = 2$ is based on the characteristics of the Poisson distribution, which determine the population's level frequency. As levels increase beyond a certain threshold, the probability mass associated with higher levels diminishes significantly, thus, as k increases, fewer agents exhibit behaviors associated with levels $k - 1$. When $k = 4$ and beyond, the behaviors of levels $k - 1$ and k become indistinguishable, leading to behavioral convergence. This truncation strategy also aligns with the nature of the Poisson-CH model and ensures that the model remains practical and interpretable within the given framework.

We demonstrate an example of how we use a Poisson distribution to define the level composition in Table 2. Given a Poisson(1.5) used to define the level distribution of N drivers, we first draw a sample of size N (i.e., counts in row 1 and percentages in row 2) and then normalize the sampled L0-L2s counts to create a 30-40-30 split (i.e., row 3) after rounding the proportions to the nearest tenth to simplify participants' reasoning. This pseudo-Poisson distribution including

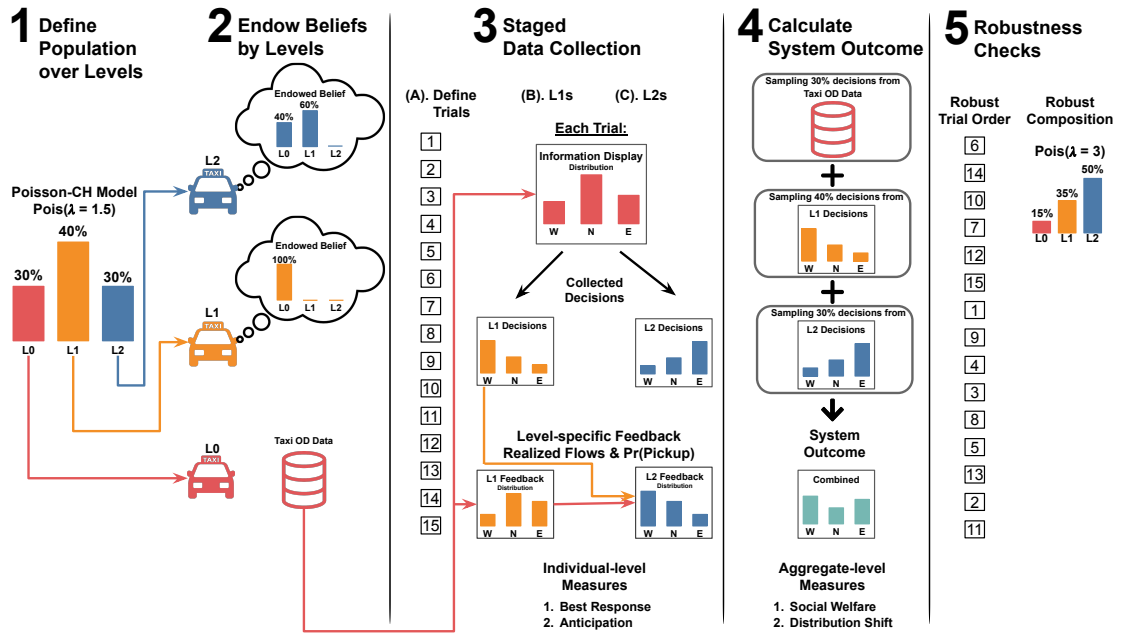

Fig. 1. Diagram of key features of our experimental design. (1) We define a pseudo-Poisson distribution using a Poisson(1.5), which includes L0s-L2s. (2) We endow level-specific beliefs by normalizing the pseudo-Poisson distribution for L1s and L2s who are our study participants. (3) We conduct a staged data collection: in each trial, participants use the information display (shown in Figure 3) to make decisions and then review level-specific feedback (shown in Figure 4). L1s and L2s complete all trials in the same order, but L1s complete the study before L2s, so that L1s' responses can be used with that of L0s (i.e., the taxi data) to construct level-specific feedback that aligns with L2s' endowed belief. (4) We calculate the system outcome by combining decisions from L0s (i.e., the taxi data) and L1-L2s (i.e., the collected responses) according to the mixture we used to define the level distribution (step 1). (5) We conduct two replications of the experiment by varying trial order and re-define the level mixture of L0s-L2s using a Poisson(3).

L0-L2 drivers represents the "true" population mixture over levels, which is *exclusive* knowledge of the principal used to aggregate the system outcomes by combining decisions using data and the collected responses from our participants.

Following the level definitions of the Poisson-CH model, our study participants are utility-maximizing but myopic agents whose decision-making is governed by the depth of thinking we endow. As illustrated in Figure 1, the non-strategic L0 players in our population are real drivers queried from the taxi data: players who made decisions based on their prior driving experience without using the shared information displays. L1 and L2 players are study participants for whom we endow levels on all study screens that present shared information displays and elicit responses. Specifically, L1 participants are told:

> *All other drivers will NOT consult the display and do NOT know the exact number of competitors in the region. They will drive according to their past driving experiences, which the taxi company has used to create the information display.*

We inform L2s that the other drivers are a mix of L0s and L1s, derived by re-normalizing the pseudo-Poisson distribution used to define all levels, which creates a 40-60 split of L0-L1. Specifically, L2 participants are told:

> *[XX]% ([level-0 count]) of drivers do NOT use the display and do NOT know the exact number of competitors in the region. These uninformed drivers will make decisions according to their past driving experience.*

[YY]% ([level-1 count]) of drivers consult the same display as you do, but each of them falsely assumes they are the only person using the display.

We evaluate individual-level decisions of participants using *level-specific outcomes* that align with the endowed belief, which we elaborate on in Section 3.2.4.

Table 2. We use a Poisson(1.5) to define all participants' level composition for trials of our main experiment using the minimum total drivers of the 15 trial weekdays, which is 598. The Poisson(1.5) generates a frequency distribution that includes 149 (25%) L0s, 184 (30%) L1s, and 139 (23%) L2s. Notice that this proportion does not sum to 100% because there are higher levels in the sample, which we omit. We then use the counts of L0-L2s (row 1) to re-normalize the distribution and round to the nearest tenth (row 3). Based on this 30-40-30 split from the pseudo-Poisson, the true level composition for this trial consisting of 598 participants is 180 (30%), 240 (40%), and 180 (30%) L0s, L1s, and L2s (row 4) after rounding again to the nearest tenth to simplify level endowment.

Level-k	0	1	2
#Obs.	149	184	139
Percent	25%	30%	23%
Rescaled	**30%**	**40%**	**30%**
True Comp.	180	240	180
Recruit 50%	0	**120**	**90**

3.2.2 Tasks and Rewards. We create the decision scenarios that constitute the trials of our experiment by sampling 15 unique but homogeneous weekdays from the taxi data used to train the counterfactual model. Participants in the experiment must repeatedly decide "where should I find my next pickup at 9 AM?" based on information displays that render both the deduced search flows and the predicted pickup probabilities, assuming all drivers act according to their past driving experiences. Although our decision scenarios have the same time setting, the search-pickup dynamics presented in the information displays vary, because each reflects decisions of a unique set of N drivers on that specific historical weekday (see Appendix A.2). We keep the time of the decision scenarios fixed at 9 AM on weekdays so that participants in the experiment are in a position to learn from repeated decisions, similar to how a taxi or rideshare driver might form a mental model of what dynamics to expect during a given driving time frame.

In each trial, we elicit participants' decisions and their corresponding anticipation of other participants' actions. Participants are asked to (1) *decide* where to search by selecting a district and (2) *anticipate* what other drivers will do by entering the number of drivers that they think will search in each district according to the endowed level. As shown by the elicitation interface in Figure 2, participants select from multiple choice options and use a dynamic input form based on previous work on eliciting Dirichlet distributions [9, 50] to provide anticipated flows. Each participant receives a base pay of $2 and a bonus of $0.2 for each trial in which their selected strategy resulted in a pickup.

3.2.3 Generating Shared Information Displays. For each trial, the information display presents flows going into the three districts by deducing decisions using the search preferences of real taxi drivers (i.e., L0s) who are involved in the decision scenario. These deduced flows are then used to predict pickup probabilities with the counterfactual model. As decisions based on real drivers' search preferences are subject to uncertainty, we use simulations to create a distribution of hypothetical outcomes. Each outcome is used as a frame for the animated display, and they are aggregated to produce the static display.

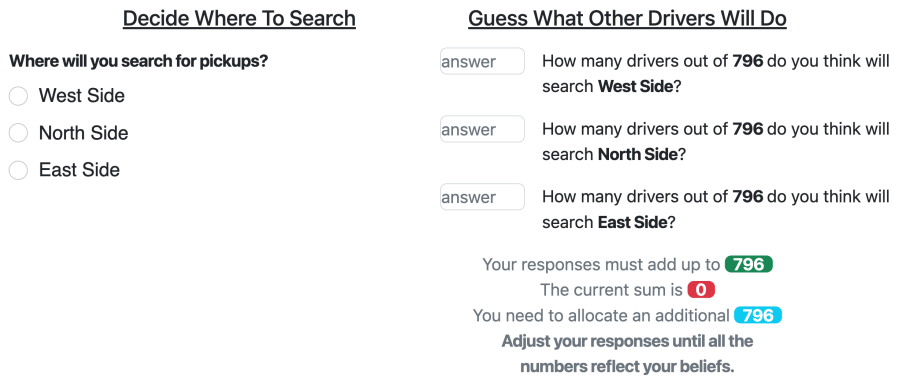

Fig. 2. The interface used to collect participants' decisions. When providing anticipation, after a participant provides guesses for two districts, the interface imputes the flow of the last district to ensure proper summation to the total number of drivers of the decision scenario. The interface dynamically updates both the current flow sum and the amount of flow to be allocated or removed if the elicited flows does not sum to the correct total.

Simulated Flows and Predicted Payoffs. We first identify a set of N candidate drivers involved in a decision scenario who would be able to search our districts at 9 AM from their current location (see *trace dyad* in Appendix A.3). We then consult each candidate driver's conditional search prior, which is encapsulated in the driver's *search dyad* $V \xrightarrow{N} S$, where V is the drop-off district of the previous trip, S is the pickup district of the consecutive trip, and the weight N is the number of occurrences of the pickup pattern (see *search dyad* in Appendix A.3). Because a *search dyad* summarizes a driver's search preferences from the current location based on her pickup history over the past ten days, we deduce a driver's search decision by sampling s using n as weight. We combine sampled decisions to deduce the flow of each district and address uncertainty by replicating this procedure 1,000 times, so each district has a distribution of simulated hypothetical outcomes including the deduced flow and the resulting pickup probabilities predicted by the counterfactual model. We provide detailed descriptions of our counterfactual model in Appendix A.4.

Visualizing Predictions as Networks. The action set (i.e., possible choices) of our congestion game contain three districts forming a traffic network describing flows. We present these predictions in the form of node-link diagrams of an egocentric network, which is a common representation of flow data. Our design choice is governed by the fact that the information display communicates two important forms of payoff-relevant information to participants: (1) deduced flows and (2) predicted pickup probabilities. Nodes representing districts are labeled with district names[2] with the corresponding predicted pickup probability as text above the node. This value is also encoded as node size and hue. Links connecting the ego and nodes of districts (i.e., alters) represent deduced flows, where the amount of flow is encoded by edge width. We position the nodes on the display to resemble an outgoing star [43] using a force-directed layout algorithm [18].

We randomly vary the display types between participants. As shown in Figure 3, some participants are assigned Network Hypothetical Outcome Plots (NetHOPs) [68], which depict prediction uncertainty more saliently by showing simulated network realizations. Following suggestions from Zhang et al. [68] on how to tune NetHOPs' visualization parameters to effectively support node-attribute and edge-attribute tasks, we render 1,000 hypothetical outcomes in a

[2]Each district of the action set represents one of three Chicago Community Areas. North Loop as "North District", West Loop as "West District", the Loop as "East District". See more detailed description in Appendix A.1.

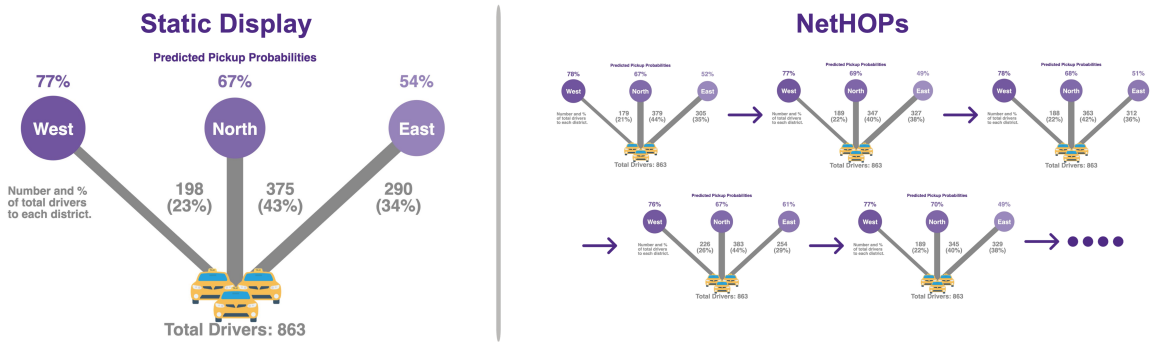

Fig. 3. Examples of the information displays used for decision scenarios varied by uncertainty quantification. **Right**: NetHOPs, which render 1,000 hypothetical outcomes, are presented in a looping animation with an animation speed of 0.2 seconds per frame using a fixed force-directed layout (i.e. anchoring $\alpha = 1$). This approach follows suggestions from Zhang et al. [68] that best support node-attribute and link-attribute tasks. **Left**: Point estimates where the rendered payoffs are the weighted averages of our simulations.

looping animation with an animation speed of 0.2 seconds per frame using a fixed layout (i.e. anchoring $\alpha = 1$). The remaining participants are assigned a static node-link diagram of point estimates, in which the deduced flows and predicted pickup probabilities visualized are the weighted averages over the 1,000 hypothetical outcomes.

3.2.4 *Evaluating Decisions and Generating Level-Specific Outcome.*

Staged Experiment Design. Our desire to incorporate both realism and control into the strategic setting, especially by introducing variation in players' levels of strategic sophistication, presents a challenge: to evaluate decisions by higher-level participants (i.e., L2s), we need access to decisions from the lower levels. While this is trivial for L0s, whose decisions are drawn from the taxi data as shown in Figure 1, to score and give feedback to L2 decisions, we need access to L1 decisions. Consequently, we designed a staged data collection procedure in which L1 participants first complete the series of trials, and are scored according to L1's endowed beliefs, then L2s complete the same series of trials, but are scored according to L2's endowed beliefs.

There are several implications of this experimental design choice. Because the post hoc decision feedback participants receive is *level-specific*, we can study individual-level performance under the assumption that players' beliefs are fixed and contrast the impact of different interface manipulations on players of different levels of sophistication. By combining L0, L1 and L2 according to the Poisson distribution used to define the frequency of all levels, we can also calculate aggregate-level system outcomes under the assumption that players are myopic (i.e., unaware of other players at the same level as them and above). Note that this requires a fixed trial order where all agents experience the scenarios in the same order, so that the aggregate results are not confounded by differences stemming from prior decision-making. The primary limitation of the staged design is that we cannot know to what extent receiving decision feedback based on the "true" population mixture including players of all levels would change the patterns of behavior that we observed.

Computing Level-specific Feedback. We generate *level-specific outcomes* that align with participants' endowed levels to score their decisions and provide feedback in each decision scenario. Recall that an L1 player believes herself to be the most sophisticated player and assumes that her competitors are all L0s. Therefore, L1s' decisions are evaluated by a *level-specific outcome* created by combining the decisions of L0s, which is exactly the distribution of search flow and the corresponding pickup probabilities we used to fit the counterfactual model. Similarly, since an L2 player believes that

she is playing against a combination of L0s and L1s, we generate L2s' *level-specific outcome* by sampling the decisions of L0s (from the taxi data) and L1s according to the proportion of L0s and L1s endowed to L2s (see Section 3.2.1), which is possible as a result of the staged data collection. To reduce sampling error, we repeat the sampling procedure 1,000 times and summarize the samples by computing the expected flow distribution. We use this distribution to predict each district's pickup probabilities by the counterfactual model and to evaluate L2's decisions.

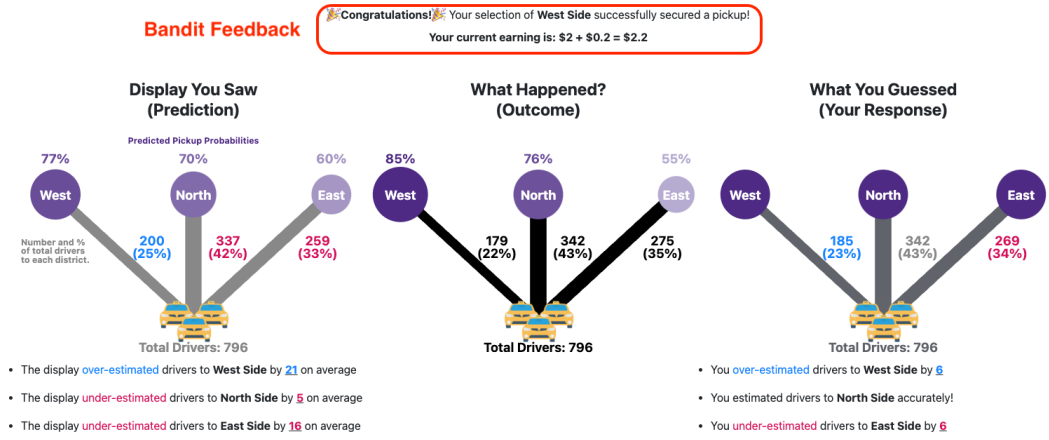

Fig. 4. Example of bandit and full feedback provided after a participant submits their responses. Both bandit and full feedback show the decision result and the current compensation for the study (i.e., highlighted in the red rectangle), which is also the only feedback information provided to participants by the bandit feedback. Participants in the full feedback condition will see three additional visualizations reminding them of the prediction they used (left), the *level-specific outcome* (middle), and their anticipation submitted with decision (right).

Feedback Structure. We evaluate two types of decision feedback: *bandit* and *full*. In the full feedback conditions, participants are presented with the *level-specific outcome* along with a reproduction of the predictions and their anticipated flows, as illustrated in Figure 4. To enhance the visibility of realized prediction and anticipation errors, we use blue color to denote over-estimation and red color to denote under-estimation in the labels of the feedback displays and the accompanying summarizing text. In the bandit feedback condition, participants receive a subset of the feedback information provided in the full feedback condition, focusing solely on the decision results, as highlighted in Figure 4.

3.2.5 Experiment Procedure. Participants are directed to our study interface[3] by Prolific. They first see a welcome page with the estimated study completion time and compensation. Participants are instructed to complete the study in a single session using Google Chrome on a large screen device. If a participant agrees to the terms by clicking a button, they are randomly assigned to one of the four treatment conditions shown in Table 1.

Upon enrollment, participants must review four detailed instruction pages. These instructions provide a comprehensive overview of the study, encompassing the experimental setting and specific guidance on using the assigned display and feedback for decision-making. After reviewing the second instruction page, participants encounter a multiple-choice question designed to assess their understanding of the display. Participants in the static display condition are prompted to select the district with a specified pickup probability. In contrast, those in the NetHOPs condition must estimate

[3]Study interface: https://strategic-performativity.com

the pickup probability of that district based on the animation. To proceed, participants need to answer the question correctly, though they have multiple attempts available.

On the third and fourth instruction pages, participants engage in a practice trial. The task page of this practice trial includes a unique instructional question, prompting them to select the district they believe most drivers will search. Participants then proceed to the feedback page, where they receive an explanation of the reward mechanism and learn how their ongoing compensation will be updated and displayed throughout each trial. After completing all 15 trials, participants have the option to describe how they utilized the information display.

3.2.6 Robustness Checks. To ensure the robustness of our findings, we conduct two additional replications of our experiment. These replications assess variation in two specific assumptions, trial order and level composition.

(1) **Robust Trial Order**: Participants complete trials in a different but fixed order than that used in the main experiment (see columns 3 and 5 of Figure 1). This approach allows us to evaluate to what extent our results might be influenced by the specific trial order.
(2) **Robust Level Composition**: Participants complete trials in the same order as in the main experiment. However, we alter the level composition by using Poisson(3) instead of Poisson(1.5) (refer to column 4 of Figure 1). This change results in an increased proportion of more strategically sophisticated L2s within decision scenarios consisting of N drivers. By following the same procedure outlined in Section 3.2.1, we modify the proportion endowed to L2s from a 40-60 split for L0-L1 (as in the main experiment) to a 20-80 split for L0-L1. This adjustment enables us to explore how potential variations in results might manifest when the level distribution leans more towards higher strategic sophistication.

3.2.7 Participants. We recruited participants from Prolific using a gender-balanced sample and two pre-screening criteria: fluency in English and being based in the US. Given our staged experimental design and the need to collect data for two additional rounds of robustness checks (i.e., a total of six data collection arms), we excluded participants who had previously taken part in the study, ensuring that each collection arm contained unique Prolific users.

Defining Level Composition. We define the distribution over levels for our decision scenarios using a Poisson(λ) where $\lambda = 1.5$ for the main experiment and Robust Trial Order experiment, and a $\lambda = 3$ for the Robust Composition experiment. As described in Section 3.2.1 and demonstrated in Table 2, we adopt a breakdown of 30% L0s, 40% L1s, and 30% L2s for the main experiment and Robust Trial Order experiment, and a breakdown of 15% L0s, 35% L1s, and 50% L2s for the Robust Composition experiment, following the same procedure.

Number of participants to recruit. Given that each decision scenario in our experiment involves hundreds of drivers, we choose to recruit 50% of L1s and L2s based on the true composition (see row 5 of Table 2) for each between-subject treatment of the main experiment (assuming the minimum number of drivers for a scenario is 598), and 25% for the two robustness checks. We then re-sample decisions with replacement from these 50% and 25% samples to create the *level-specific outcomes* shown to L2s and the system outcomes, which are generated by combining the decisions of all three levels according to the level distribution defined for each experiment. Table 3 presents the number of participants we recruited by levels and by collection arm. Note that the total number of participants we recruit for the robust composition check is 300. This is because this collection arm follows the same trial order as the main experiment, allowing us to use L1 responses from the main experiment without additional recruiting.

Table 3. The number of participants recruited for each collection arm by size. The parentheses in each table cell include the number of recruited participants for each of four treatment conditions per level (column) and per collection arm (row).

Collection	Level-1	Level-2	Total
Main Experiment	480 (4×120)	360 (4×90)	840
Robust Trial Order	240 (4×60)	180 (4×45)	420
Robust Composition	212 (4×53)	300 (4×75)	300

Note: The 212 L1s from the Robust Composition check are sampled from those of the main experiment because L1 tasks are identical for these two collection arms.

3.2.8 Analysis Method. We pre-registered an analysis plan, which includes four dependent variables: two at the aggregate-level and two at the individual-level.

Aggregate-level Dependent Variables. In essence, a principal cares about aggregate-level outcomes like the efficiency of the system in terms of matching drivers and riders or the magnitude of the difference between what is realized and what is predicted. We calculate aggregate-level outcome variables based on the system outcomes, created by combining L0's decisions from the taxi data and L1's and L2's decisions collected from participants in each trial according to the pseudo-Poisson distribution used to determine the frequency distribution across all levels. Since the number of participants we recruit per level per trial is fewer than what is defined by the decision scenario, we sample participants' decisions with replacement to achieve the correct number needed for each decision scenario. To mitigate the effect of sampling error, we generated system outcomes 500 times per trial, forming a distribution for each aggregate-level outcome variable. It is important to note that these system outcomes are different from the *level-specific outcomes* used to evaluate participants' individual-level decisions, and can only be calculated after both L1 and L2 stages of data collection are complete.

Given these per-trial outcome distributions, we define the **social welfare ratio** for each trial as the proportion of social welfare achieved (i.e., total realized pickups) out of the total possible social welfare obtainable (i.e., maximum pickups). We use a non-linear optimization via the augmented Lagrange method [67] that employs random initialization and multiple restarts [29] to identify the optimal flow distribution for each trial. We define an objective function that takes an input flow vector of length three and outputs a numerical value representing the number of drivers who failed to get pickups given constraints that ensure the input flow vector sums to the total number of drivers in a decision scenario and that the minimum and maximum flow to a district does not exceed the bounds given by the historical data. Full details are given in the Supplemental Material.

We calculate the **distribution shift** for each trial as the Earth Mover's Distance (EMD) [56, 57] between the distribution of deduced search flows shown to participants on the display and the flows simulated from the system outcomes. EMD quantifies the discrepancy between the displayed flows and the flows of the system outcome by the minimum amount of "work" needed to match the two flow distributions. To compute EMD, we map both flow distributions into a 2D grid and find the minimum amount of Euclidean distance required to shift and align the flow distributions, which is then standardized by the total number of drivers in the decision scenario. We describe the cost function used to optimize the EMD computation in our Supplemental Materials.

Individual-level Dependent Variables. We define two individual-level response variables to help us gain insights into how participants endowed with varying levels of strategic sophistication are affected by our experimental manipulations. At the individual-level, we define **best response** as a binary indicator of whether a participant selected the district that results in the highest expected pickup probability based on the endowed level. For L1s, best response amounts to simply choosing the district with the highest expected pickup probability as shown on the display. For L2s, the best response is the district that results in the highest expected pickup probability based on the *level-specific outcome* formed by combining L0 and L1s' decisions under the endowed proportion of L2s. Evaluating the best response of L1s and L2s over repeated decisions allows us to observe how access to prediction uncertainty and realized prediction error may be more or less useful for agents acting under particular beliefs about the behaviors of other agents.

Anticipation error measures how well a participant anticipates other participants' search decisions. Evaluating anticipation error over repeated decisions allows us to observe how the provision of realized prediction error, in particular, may help participants anticipate others' behaviors even if the predictions of the information display are "inaccurate". Similar to distribution shift, we compute the EMD between each participant's anticipated flow distributions of how many drivers would go to each district and the *level-specific outcome* used to evaluate decisions under the endowed belief (see Section 3.2.4).

Statistical Models. For the two aggregate-level variables, we pre-registered regression models for welfare ratio and distribution shift, using the expected values calculated from 500 simulations as the dependent variables. We specified the maximal models [4] for the aggregate-level variables by Wilkinson-Rogers-Pinheiro-Bates syntax [52, 64] as follows:

Aggregate-level Model 1 Welfare Ratio (*ratio*)

1: $ratio \sim \text{Beta}\ (\mu, \phi)$
2: $\text{logit}(\mu) = display * feedback * trial$
3: $\log(\phi) = display * feedback * trial + maxProb$

Aggregate-level Model 2 Distribution Shift (*DS*)

1: $log(DS) \sim \text{student\_t}\ (\nu, \mu, \sigma)$
2: $\mu = display * feedback * trial$

Here, *trial* and *maxProb* are numeric, while *display* and *feedback* are categorical. For the welfare ratio model, we use a Beta distribution as the likelihood (line 1, left), parameterized by mean (μ) and precision (ϕ). We apply a logit link for μ (line 2, left) and include all experimental variables, *display*, *feedback*, and *trial*, as predictors on μ of the likelihood. We apply a log link for ϕ (line 3, left) and include an additional predictor *maxProb*, which is the maximum pickups optimized based on the system outcome to control the amount of spread for the achieved welfare. For the distribution shift model, we use a student-t distribution as the likelihood for log-transformed *DS* (line 1, right), parameterized by ν (degree of freedom), μ (mean), and σ (scale) and include all experimental variables as predictors for μ (line 2, right).

To model individual-level variables while accounting for the experimental design, we pre-registered Bayesian Linear Mixed-effect Models for best response and anticipation error. We specified separate models by levels, since the tasks that an L1 completes is different from an L2, and separate models for each collection arm, since level composition and trial order change for the two robustness checks. We specify the individual-level maximal models as follows:

Individua-level Model 1 Best Response (*BR*)

1: $BR \sim \text{Bernoulli}\ (p)$
2: $\text{logit}(p) = display * feedback * trial + (trial \mid ID)$

Individua-level Model 2 Anticipation Error (*AE*)

1: $log(AE) \sim \text{student\_t}\ (\nu, \mu, \sigma)$
2: $\mu = display * feedback * trial + (trial \mid ID)$

Here, *trial* is numeric, while *display*, *feedback*, and *ID* are categorical, with *ID* representing each participant's unique Prolific ID. Line 1 of both models presents the assumed likelihood functions. Since best response is binary, we use a Bernoulli distribution as the likelihood, with parameter p indicating the best response rate. To account for outliers in participants' anticipation, we use a Student-t distribution as the likelihood for log-transformed *AE*, parameterized by ν (degree of freedom), μ (mean), and σ (scale). Line 2 presents the hierarchical linear models, for which we use a logit link for best response rate p. Both individual-level models estimate fixed effects of trial, display type (static or NetHOPs), and feedback structure (bandit or full). Because participants may have varying baseline performances and demonstrate different variations in performance across trials, we specify random effects for participants as random intercepts and random slopes for effects of *trial* and the intercepts.

We used a standard Bayesian workflow [24] to check our model fits and include model diagnostics in the Supplemental Material. We report expected performance with the median point estimate of the expectation with uncertainty expressed as a 95% highest posterior density interval (i.e., credible interval (CI)) for each variable and condition, marginalizing over trials unless examining learning effects.

4 RESULTS

In total, we received 1,573 valid responses. Table 4 summarizes the number of unique participants by treatment, level, and collection arm, with these numbers highlighted in bold. We removed four participants from the total number recruited according to our pre-registered exclusion rule (i.e., two L1s of the main experiment and two L2s of the Robust Trial Order experiment).

Table 4. The number of valid responses received by condition, level, and collection arm.

Collection	Main Experiment		Robust Trial Order		Robust Composition	
Level-k	Level-1	Level-2	Level-1	Level-2	Level-1	Level-2
Valid	**480**	**361**	**240**	**181**	212	**311**
Static+Bandit	115	103	63	40	53	78
Static+Full	115	86	59	44	53	70
NetHOPs+Bandit	122	92	52	49	53	84
NetHOPs+Full	128	80	66	48	53	79

Note: The 212 L1 responses from the Robust Composition check are sampled from those of the main experiment because L1 tasks are identical for these two collection arms.

4.1 Data Preliminaries

Table 5 presents the study completion time, from which we see that the completion time depends on the endowed levels, display type, and feedback structure. Participants who were endowed with L2 beliefs, who viewed NetHOPs, and who received full feedback took more time to complete the study.

Recall our participants could optionally describe how they used the interface to make decisions. We lightly analyzed their statements by sampling 100 comments from each level across collection arms. From 84 meaningful L1 comments, we found four participants who either misunderstood the payoff information shown in the display or misinterpreted the strategic setting. From 83 meaningful L2 comments, we found that the level-endowment success rate is slightly lower (84%), with 13 participants acting like L1s or randomly guessing. Hence, the realized Poisson distribution in the main experiment might skew slightly more toward L0s and L1s than intended.

Table 5. Summary of study completion time by treatment condition.

Group	Min.	Median	Mean	SD	Max
Pooled	4.4	18.1	20.4	10.6	132
Level-1	4.4	17.4	19.4	9.5	77
Level-2	5.1	18.9	21.5	11.5	132
Static + Bandit	4.4	16.8	18.8	10	83.1
Static + Full	5.9	18.7	21.1	10.8	78.8
NetHOPs + Bandit	5.1	16.4	19.3	10.9	132
NetHOPs + Full	6.2	20.7	22.5	10.2	77

4.2 Aggregate-level Outcome

We first examine the two aggregate-level variables, welfare ratio and distribution shift, calculated from the system outcomes by combining the decisions according to the true mixture over the levels. The welfare ratio in our experimental setting captures the efficiency of the system, measured by the proportion of realized pickups relative to the maximum number of pickups that could have been attained from a decision scenario. The distribution shift quantifies the difference, in EMD, between the deduced flow shown on the display and the realized flow of the system outcome. We present the 95% CIs of the median expected outcomes of the posterior predictive distribution by treatments shown in Figure 5 (A), and examine its variations by trial in Figure 5 (B).

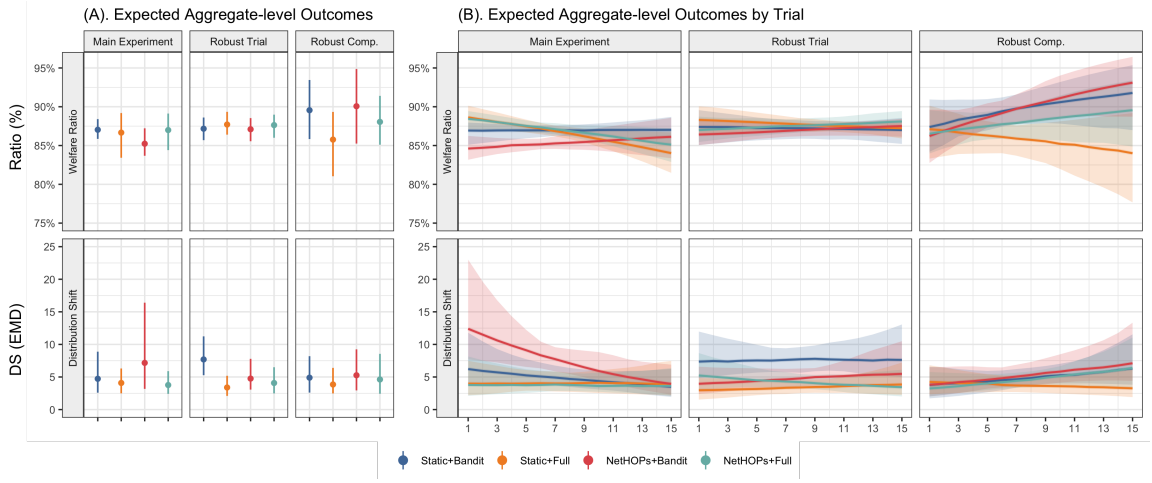

Fig. 5. (A) Median point estimates of the expected welfare ratio (top) and distribution shift (bottom) resulted from the system outcome. We expressed uncertainty as 95% credible intervals (CIs) predicted by the fixed effects of both aggregate-level models for each treatment marginalized over trials. (B) Same as (A), we present both the point estimate and the corresponding 95% CIs for welfare ratio (top) and distribution shift (bottom) for each trial to examine changes in system outcome by trial.

4.2.1 *Welfare Ratio.*

Welfare Ratio by Interface. When using a Poisson(1.5) to define a 30-40-30 mixture of levels (i.e., main experiment and Robust Trial Order), we find similar expected welfare ratios across treatment interfaces. In all treatments, the expected

ratios fall within the range of 83% to 89%. (Figure 5 (A) top-left and middle). However, when combining decisions in the Robust Composition experiment using a Poisson(3) that produces a split more skewed toward higher-levels (15-35-50), we observe a higher median expected ratio on average with greater variance across interfaces. The display type, either NetHOPs or static displays, appears to have minimal effect on welfare ratio of the system outcomes. However, bandit feedback (90%; CI[86%, 94%]) appears to produce slightly higher welfare than those produced by full feedback (87%; CI[82%, 91%]).

Trial-level Variations. We observe consistent interface effects when examining the expected welfare ratio by trial shown in row 1 of Figure 5 (B). From the main experiment, changes in ratio by trial appear to depend more on the feedback structure, with bandit feedback leading to a slight increase in welfare ratio over trials (red and blue) and full feedback leading to a slight decrease (yellow and green). The trial-level variation in the welfare ratio observed in the Robust Trial Order experiment is similar to those observed from the main experiment, except that the feedback effect is weaker, as reflected by flatter slopes. With a level distribution that skews more toward higher levels (Robust Composition experiment), although welfare ratios are similar in early trials, they begin to diverge such that all interfaces have increasing welfare with more trials except for static displays with full feedback (yellow), where welfare decreases.

In summary, we observe fairly robust welfare variations over the three experiments when it comes to: (1) static displays and full feedback resulting in a lower welfare ratio by trial, and (2) NetHOPs displays and bandit feedback leading to a higher welfare ratio over trials.

4.2.2 *Distribution Shift.*

Distribution Shift by Interface. Under a 30-40-30 mixture of levels from a Poisson(1.5) (main experiment and Robust Trial Order), we find full feedback leads to lower and less extreme expected distribution shift than bandit feedback on average (e.g., main experiment, Full 3.9; CI[2.4, 6.3] and Bandit 5.7; CI[2.6, 14.1]). As shown in Figure 5 (A) bottom-left, this effect is especially prominent for NetHOPs where variance in the expected amount of distribution shift is quite high with bandit feedback (red). We see a smaller version of this effect in the Robust Composition experiment. Static displays may produce slightly lower and less extreme expected distribution shift than NetHOPs (e.g., main experiment, Static: 4.4; CI[2.5, 8.1] and NetHOPs: 4.7; CI[2.5, 14.1]) but the estimates have high uncertainty.

Trial-level Variations. From the main experiment (Figure 5 (B) bottom-left), changes in distribution shift over trials appear to be driven by the feedback structure. With the bandit feedback, distribution shift decreases with more trials. With full feedback, distribution shift remains fairly consistent with more trials. Although distribution shift varies substantially depending on the interface in the early trials, the expected distribution shift appears to converge by the last trial across the interfaces. However, we observe a different pattern for distribution shift in the Robust Trial Order experiment (Figure 5 (B) bottom-middle), where changes in distribution shift over trials are more similar, as reflected by relatively flat slopes. For the Robust Composition experiment (Figure 5 (B) bottom-right), expected distribution shift tends to increase by trials across interfaces, except for the condition of a static display with full feedback. Although the changes in distribution shift vary by trial order and level composition, a static display with full feedback can consistently produce slightly lower distribution shift across the experiments.

4.3 Individual-level Analysis

We analyzed participants' individual-level best response rate and anticipation error to explore the potential causes of the observed variations in the aggregate-level variables. Recall that in our experimental setting, selecting a best

response means that a participant decides to search a district that has the best chance of resulting in a pickup based on the level-specific outcome, while the anticipation error captures the difference, in EMD, between the anticipated flow and the flow of the level-specific outcome. We report both individual-level response variables using the 95% CIs of the median expected value from the posterior predictive distribution by treatment conditions and examine them by trial.

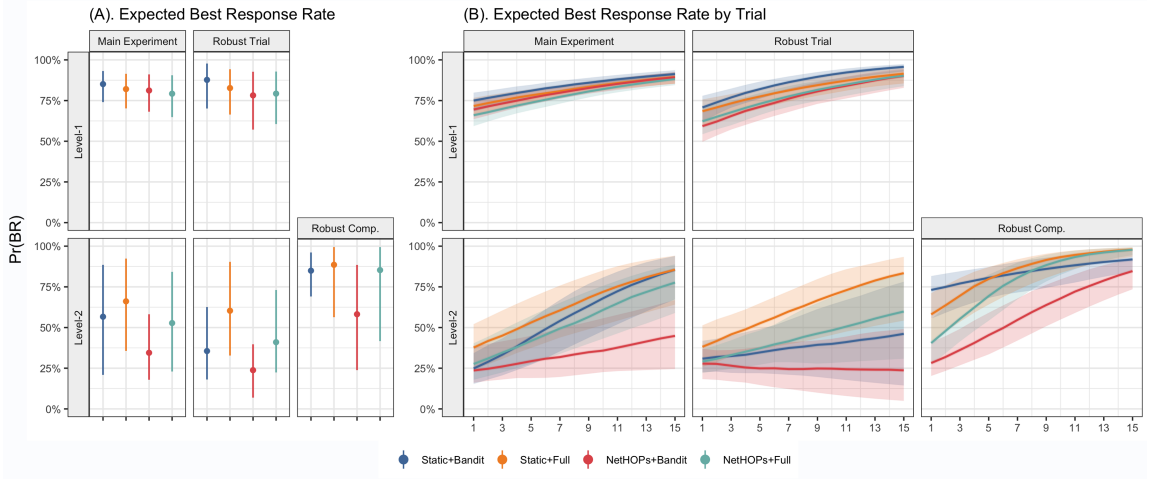

Fig. 6. (A) The median point estimates of the expected best response rate with uncertainty expressed as 95% credible intervals (CIs) predicted by the fixed effects and marginalizing over trials. (B) As in (A), we present both the point estimate and the corresponding 95% CIs of the expected BR rate for each trial to examine changes by trial. ***Notice we omit L1s of robust composition check because these L1s' responses are sampled from those of the main experiment.***

4.3.1 Level-1 Best Response. We observe minimal differences between L1's best response rate across treatments from both the main experiment and Robust Trial Order experiment, shown in the top-left and middle of Figure 6 (A). The high best response rates of L1s are to be expected: because L1s are endowed with the belief that the displayed prediction is "accurate", their best response is to choose the district that shows the highest predicted pickup probability from the display. In both experiments, we find (1) static displays lead to slightly higher BR rates for L1s than NetHOPs (e.g., main experiment, Static: 84%; CI[71%, 93%] and NetHOPs: 80%; CI[66%, 91%]), and (2) bandit feedback leads to slightly higher BR rates over full feedback (e.g., main experiment, Bandit: 83%; CI[70%, 93%] and Full: 80%; CI[67%, 92%]). It may be that when best responding to a display that is straightforward, processing prediction uncertainty and viewing full feedback can make some participants second-guess how to best respond. However, all the interface effects observed are quite small for L1s.

We observe slight but consistent improvements in BR rates in all treatments for L1s as they completed more trials of the main experiment and Robust Trial Order experiment (Figure 6 (B) top-left and middle). We can see that L1s who used static displays and bandit feedback (blue) are more likely to best respond than those who used the alternatives.

4.3.2 Level-2 Best Response. L2s' best response rates (Figure 6 (A) row 2) are much lower than those of L1s across treatments, with greater uncertainty in the estimates. This is likely because L2s face a harder task: they must reason about the distribution of other players' actions using a display that is not expected to be "accurate". The larger variance that we observe is partially due to high trial-level variance, as some L2s appear to have been able to considerably improve their best response rates with repeated decisions.

At a high level, we find that (1) L2 participants who received full feedback were slightly more likely to best respond than those who received bandit feedback (e.g., main experiment: Full 59%; CI[26%, 90%] and Bandit 40%; CI[19%, 85%]), and (2) L2s who viewed static displays were more likely to best respond than those who viewed NetHOPs (e.g., main experiment: Static 61%; CI[25%, 91%] and NetHOPs 40%; CI[18%, 80%]). These effects persist in the Robust Trial Order and Robust Composition experiments, though intervals overlap. The use of static displays and full feedback may be the most effective for encouraging L2s to best respond. Our results provide weak evidence that the emphasis on prediction uncertainty makes it harder for L2s to anticipate how others will respond to a display, which Kayongo et al. [38] had speculated but were unable to validate from their study. However, the high uncertainty in the estimated L2 performance, likely due to the difficulty of best responding to a display under beliefs about a mixture of other agents over levels, prevents us from drawing any strong conclusions.

One other difference we observe is that the best response rates for L2s are noticeably higher across treatments in the Robust Composition experiment. We expect this is because more L1s in the mixture over levels that L2s are endowed (i.e., 80% L1s in Robust Composition vs. 60% L1s in main experiment) may make it easier for L2s to anticipate others' actions. With more lower-level players in the mixture of belief that L2s are endowed, more competitors will choose to search in the most lucrative district following the display.

When examining L2s' best response rate by trials (Figure 6 (B), row 2), we find steeper increasing slopes than those of L1s, providing evidence that L2s can over time learn a mapping between predicted and level-specific outcomes. We see consistent treatment effects where static displays and full feedback tend to help L2s best respond over trials. We observe that L2s who use a "bad" display design for their task (namely NetHOPs and bandit feedback) may cease to learn at all depending on the order of decision scenarios (Robust Trial Order experiment in Figure 6, bottom-middle).

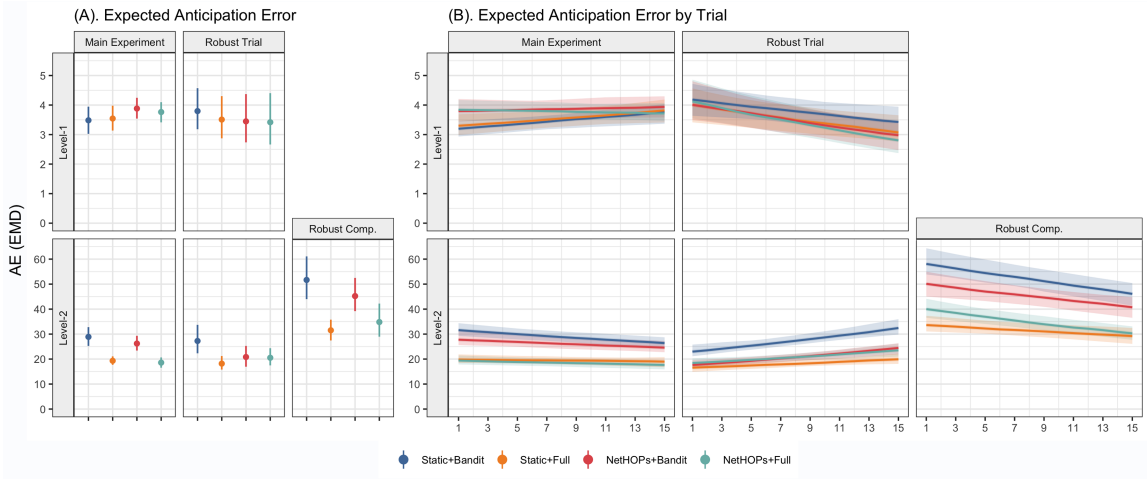

Fig. 7. (A) Median point estimates of the expected anticipation error in EMD with uncertainty expressed as 95% credible intervals (CIs), accounting for all fixed effects and marginalizing over trials. EMD describes the averaged Euclidean distance a participant's anticipated flow distribution must move in order to match the flows of the level-specific outcome. (B) Same as (A), we present both the point estimate and the corresponding 95% CIs of the expected anticipation error for each trial to examine changes in performance by trial. ***Notice the y-axis scales are different between L1s (top) and L2s (bottom), and we omit L1s of robust composition check because these L1s' responses are sampled from those of the main experiment***.

4.3.3 Level-1 Anticipation Error. Between our main experiment and Robust Trial Order experiment, the amount of anticipation error for L1s, with EMD measured in units of Euclidean distance, is quite small, with the CIs closely overlapping between a range of 2.7 and 4.5 (Figure 7, row 1). When examining the variation in L1s' anticipation error by trials (Figure 7, row 1), although the signs of the slopes are different, the slopes are all small, indicating minimal improvement or deterioration in L1s' ability to anticipate level-specific outcomes.

4.3.4 Level-2 Anticipation Error. For L2s, the expected anticipation error is much higher in all treatments compared to that of L1s. As shown in Figure 7 (A) row 2, we observe strong feedback effects across all three experiments, where L2s who received full feedback show a clear advantage over those who received bandit feedback in anticipating competitors' actions (e.g., main experiment, Full: 19; CI[16.9, 21] and Bandit: 27.3; CI[23.8, 32.2]). We observe no clear effect of using NetHOPs versus static displays.

We observe higher anticipation error for L2s in the Robust Composition experiment. This may seem to contradict the results of L2s' best response rate observed from this experiment, which were higher than those from the other experiments, presumably because of the higher proportion of L1s they attempted to best respond to. We speculate that anticipation error is a more fine-grained measure. Estimating how many other players will choose to search in each district may be difficult, even if the best response district is obvious to an L2.

Examining L2s' anticipation error by trial (Figure 7, row 2) indicates that full feedback can consistently help participants more accurately anticipate others' decisions across trials in all three experiments. The results of the Robust Trial Order experiment suggest that variations of anticipation error over repeated decisions can depend on the specific sequence of prior decision scenarios, as anticipation error decreases in the main experiment and Robust Composition experiment and increases in the Robust Trial Order experiment.

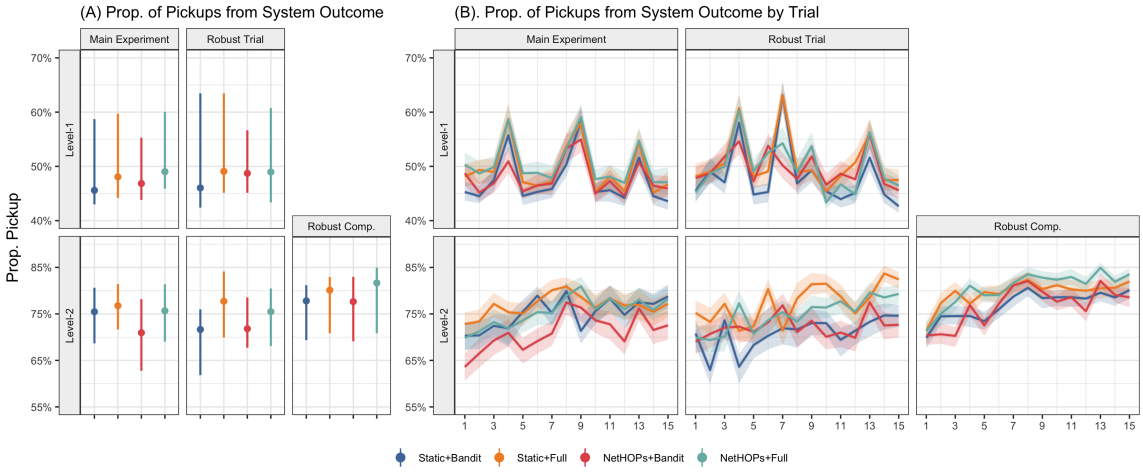

Fig. 8. (A) The median proportion of L1s and L2s who secured pickups from their selected district with 95th percentile intervals (PI) summarizes the distribution of pickup proportions from 500 simulations of the system outcomes, marginalized over trials. (B) As in (A), we present both the median and 95% PI of L1 and L2s' pickup proportion for each trial to examine changes by trial. ***Notice the y-axis scales are different between L1s (top) and L2s (bottom), and we omit L1s of robust composition check because these L1s' responses are sampled from those of the main experiment***.

4.4 Connecting Individual-level Response and Aggregate-level Outcomes

We analyzed the individual-level best response rate and anticipation error against the level-specific outcomes. A question naturally arises: do we observe the same or similar display or feedback effects when participants' decisions are evaluated against the system outcome that is realized by combining the decisions of L0s, L1s, and L2s? We construct 95th percentile intervals (PI) using the median as point estimate to evaluate the proportion of L1s and L2s who secured pickups from their selected district based on 500 simulations of the system outcomes for each type of interface, marginalized over all trials in Figure 8 (A) and by individual trial in Figure 8 (B).

From our simulations, the PIs for L1s' proportion of pickups (Figure 8 (A), row 1) suggest a strong right-skewness, with most simulated observations falling below or close to 50%, indicating that the chance of pickup for L1s, in general, is generally low and accompanied by high uncertainty. When marginalizing the proportion of pickups for L1s using all simulations across the three experiments, the display and feedback effects for L1s are minimal: the difference in proportion between Static display and NetHOPs is -0.6% (PI[-6%, 7%]), while that between full and bandit feedback is 1.7% (PI[-4%, 7%]).

The noticeable difference between the proportions of pickups for L1s and L2s (row 1 and row 2 of Figure 8 (A)) is that L2s have a clearly higher chance of getting pickups due to their strategic sophistication. Similar to the results when evaluating L2s' best response rate under their level-specific outcomes, we find minimal differences between using a static versus NetHOPs display for L2s' pickup probability in the system outcome (0.8%; PI[-6.9%, 8.6%]). We again see a higher proportion of pickup when L2s use full feedback (3.4%, PI[-3%, 10.4%]). In summary, the interface effects identified from our individual-level results appear to hold when evaluating decision-making against the system outcome as reflected by each level's proportion of pickups.

4.4.1 Level-k Decisions and Welfare Ratio. The observed trial-level variations in the social welfare ratio can be contextualized by several aspects of the individual-level decision. Design choices, such as displaying prediction uncertainty and error, have limited impact on L1s who respond to an "accurate" display. Due to the simplicity of their tasks, L1s can easily identify their best response district (e.g., West) from the information display. In contrast, although L2s face more challenging tasks, they can improve their decision-making when using an interface with static displays and full feedback. While the best response rate of L2s is generally lower than that of L1s, as some L2s may behave similarly to L1s due to their incomprehension of the endowed level, the interface's provision of realized prediction error (i.e., full feedback) can help L2s to identify their best response district (e.g., East) over repeated decisions. Consequently, as participants complete more trials, more L2s who initially misunderstood their level endowment and behaved like L1s can recognize the opportunity to profitably deviate from L1s' best response district.

It is important to note that each district of our congestion game has a limited capacity of pickups (see Figure 11 in Appendix A.4). While L2s deviating to their best response district can increase their overall share of pickups, it also decreases the probability of pickup in that district until the remaining district (e.g., North), which is not the best response for either L1s or L2s, becomes a new opportunity of profitable deviation. However, this opportunity goes unnoticed by both L1s and L2s who are myopic, resulting in a reduction in the social welfare ratio achieved by the system outcome and suggesting that design choices that can support individual-level decisions for certain groups can lead to a worse collective system outcome.

4.4.2 Level-k Decisions and Distribution Shift. The quantification of distribution shift using EMD is useful in describing the overall difference between flows shown on the display and those observed in the system outcome. However, the

EMD computation does not consider flows at the district level, specifically the difference between the deduced number and the actual number of drivers searching in each district. Because flows to each district are driven by individual-level decisions, the amount of distribution shift that arises from a system outcome is naturally associated with the number of participants who were able to select the best response strategy based on their endowed levels. Additionally, we observe that the distribution shift is influenced by interface design factors, that is, an interface with full feedback has a robust feedback effect, producing system outcomes with lower distribution shift than its bandit counterpart. We performed an exploratory analysis, investigating the trial-level variations of distribution shift through the lens of the flow proportion to each district under the influence of the two feedback structures. Our goal is to answer the high-level question: how does the feedback structure influence the flows to each district, prompting changes to distribution shift over trials?

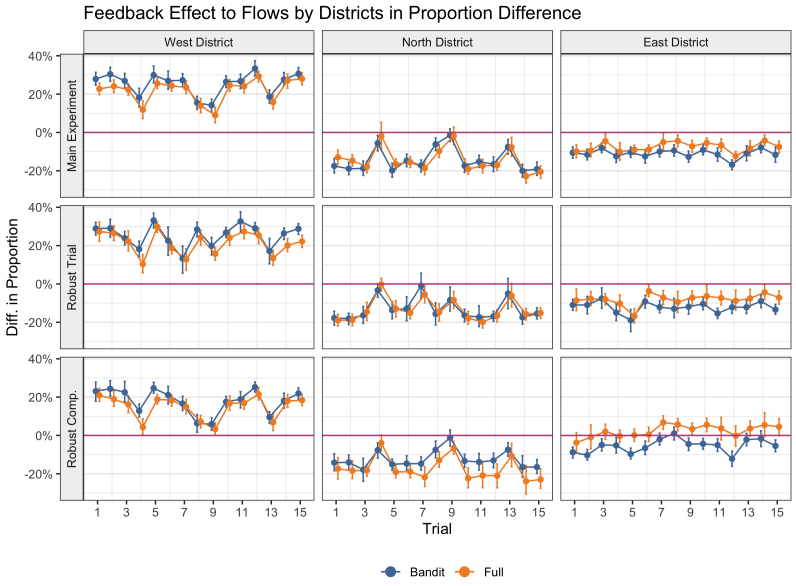

Fig. 9. The median proportion of drivers who searched in each district relative to the deduced flow shown in the information display. The intervals representing 95th percentile summarize district flows from 500 simulations. The red horizontal line centered at 0 can be viewed as a benchmark so that a positive proportion difference means more drivers go to a district, while a negative proportion difference means fewer drivers go to a district. Colors differentiate interfaces with full feedback from those without.

In Figure 9, we dissect flows of the system outcome of each trial by district (i.e., in West, North, and East) for each collection arm. We then present the differences between the observed flows for each district relative to those shown on the information display. Recall system outcomes are created by simulation. Therefore, we create 95th percentile intervals that summarize the difference in proportion going into each district that summarizes simulations. When inspecting the differences in proportion corresponding to each district across trials, we see a consistent pattern, in which more participants choose to search in the West district (i.e., above the red horizontal line centered at 0% in Figure 9, left) and fewer participants choose to search in the East or North districts, compared with the flows shown in the display. This behavior is expected because (1) choosing the West District is the best response strategy under L1s' endowed beliefs (see Table 6 of Appendix B), and (2) level endowment may fail for some L2s, who instead choose to search in the West district because it is the most lucrative location shown in the display.

When inspecting the influence of receiving full feedback (yellow intervals in Figure 9) on the difference in flow proportion to each district, we see that relatively more participants (i.e., including both L1s and L2s) choose to search in the East district (L2s' best response strategy), while fewer searched in the West district (L1s' best response strategy). Recall our individual-level results on best response rate suggested that (1) L1s tend to have relatively high and increasing best response rates across trials and (2) full feedback helps L2s with best responding. Therefore, the gaps that represent the difference in flow proportions by feedback structure observed from the West and East districts suggest that as participants complete more trials, more L2s are able to identify their best response strategy corresponding to their endowed level (see Figure 6 (B), row 2, yellow). As more L2s choose to profitably deviate and L1s maintain a high BR rate, the distribution shift of the system outcome may reduce. Hence, our results imply that a principal can expect reduced distribution shift by tailoring the interface to the needs of more strategically sophisticated players within a repeated strategic decision-making context. As the amount of distribution shift gets lower with repeated decisions, predictions shown on the information display become more "accurate" retrospectively relative to the system outcome and hence improve trust and reliance on the information display.

5 DISCUSSION

The results of our experiments highlight the impact of design choices, specifically the provision of prediction uncertainty and error, on decision-making among agents with varying levels of strategic sophistication. Methodologically, we demonstrate how level endowment can be useful for understanding the dynamics of system outcomes that are otherwise difficult to disentangle. We introduce a staged experimental design as an alternative to synchronous experiments, particularly useful in situations where collecting decisions from a large number of participants simultaneously is not feasible. Moreover, we emphasize the value of Cognitive Hierarchy Models as a powerful tool for researchers studying the design of information display for strategic settings that is naturally aligned with calls for behavioral researchers to take effect heterogeneity more seriously [7, 23].

More specifically, our results suggest that the more strategic the user population of the prediction, the more important it becomes to design information displays that are effective in anticipating how other agents will respond to the predictions. One way to facilitate anticipation according to our results is to use a static display, rather than emphasizing uncertainty. More importantly, when the prediction of an information display may be seemingly inaccurate in hindsight as a result of agents' strategic response, providing information on the realized prediction error can help agents make more informed decisions despite distribution shift. This result has implications for the study of trust in data-driven predictions. Prior work [42] suggest users' trust in predictive model outputs is affected by the observed accuracy of the model. When real-world decision-makers perceive a gap between the predicted and the realized outcomes, they may lose trust and stop relying on the predictions, rather than trying to infer how they are wrong and adjust their decision strategies. However, in a strategic setting like we study, a consistently "wrong" information display that is transparent about its error can be more useful to agents than one that is not forthright about its error. An interesting pursuit for future work is to study agents' trust in such settings. For example, when access to an information display or post hoc decision feedback carries a cost, it raises the question of what factors influence agents' willingness to tolerate an inaccurate display that can potentially improve their decision-making and how agents can effectively assess performance gains in the presence of an incorrect display.

Another important implication of our results is that information displays that help more sophisticated agents better anticipate others' responses can lead to a differential advantage that results in less total welfare across the population, but, at the same time, also reduces distribution shift. While providing an interface that can communicate realized

prediction errors to the entire *user population* appears to be the best default to maximize utility at the individual level, we find that chasing this objective will not necessarily result in greater overall social welfare. In fact, interfaces that did not disclose realized prediction error were better at maximizing the total possible social welfare of the system. However, while the deployment of an interface that visualizes realized prediction errors may have a negative impact on the system's social welfare, a principal can anticipate a reduction in distribution shift. This reduction implies a decrease in the disparity between the predicted outcome and the realized outcomes, improving the perceived accuracy of the information display for agents after decision-making. As a result, it improves agents' trust and reliance on the information display in a repeated strategic decision-making context.

Our findings suggest that there are not only ethical considerations that arise in negotiating the trade-off between individual-level utility and aggregate-level social welfare, but also an important trade-off between social welfare and users' potential trust and reliance of the information display.

5.1 Challenges of Estimating Display Effects in Strategic Settings

We study a situation where the predictions from an information display is based on historical data that is unlikely to be a good reflection of the behavior that results when agents use the display. Rather than trying to reduce the disparity between the predictions and the realized outcome, our interest is in how different design features of an information display impact individual-level decision-making and aggregate-level system outcomes and their stability over repeated decisions over time. Because the relationship between these two types of measures can be complex, studies like ours take on a challenging problem.

We believe that our goals are an important complement to approaches that aim to "prescribe" a desirable aggregate-level system outcome through a display. Such approaches include information design, which utilizes selective information disclosure (e.g., [15]); the visualization equilibrium [38] where the displayed prediction matches the realized outcome, essentially represented by a fixed point of the agents' behavior and the visualization function that produces the visualization; and theoretical solutions in machine learning research for addressing distribution shifts induced by predictive algorithms in most social prediction context (i.e., performative prediction), which involves defining stable or optimal fixed points of retraining against expected future outcomes (e.g., [46, 47, 51]). Attempts to inform design with knowledge about causal effects of display characteristics, as we pursue, tend to require fewer assumptions. However, this does not mean that identifying such effects experimentally is easy. The small number of trials we conducted imposes limitations on the dynamics we can observe. Future work might attempt to observe behavior over a longer duration, to understand the patterns of performance improvement and convergence across interfaces with extended usage

5.2 Presentation of Prediction Stimuli

In our experiment, we utilized graphs to present prediction stimuli and employed NetHOPs as a counterpart for more salient uncertainty quantification. This design choice was partially motivated by the structure of our congestion game, which comprises three districts forming a traffic network that can be most naturally visualized using a graph. As detailed in Section 3.2.3, our presentation is efficient in conveying important payoff-relevant information to participants in a clear and compact manner, thereby better streamlining the decision-making process.

However, we acknowledge that there are alternative visualization methods that can be employed. For instance, the same prediction stimuli can be illustrated using clustered or side-by-side bar charts in a small multiple form with uncertainty quantified by conventional confidence intervals. Our decision to use a frequency-based uncertainty visualization technique like NetHOPs stems from their ability to emphasize uncertainty to such an extent that participants

are compelled to consider it, rather than merely focusing on point estimates or other common biases and misconceptions that can arise when using conventional confidence intervals [5, 14, 27]. Still, we think it is beneficial for future research to compare performance of decision-making using alternative visualization techniques to present prediction stimuli and quantify the associated uncertainty for a more comprehensive understanding of the robustness and generalizability of our findings.

5.3 Limitations and Future Work

Our proposed staged experimental design could be applied to most non-cooperative games under the level-k framework, including games of different payoff structures, asymmetric information, and involving dynamic environments. However, this design is constrained by the level-k frameworks in that we can only evaluate agents' decisions using a level-specific outcome that aligns with the endowed belief (see section 3.2.4). Our design reinforces level endowment by making players myopic, which leads to consequence that we were unable, with this setup, to leverage players' intrinsic strategic sophistication and investigate the dynamics of decision-making and system outcome when players' levels of sophistication can change over repeated games in light of the results of previous plays. Future work might attempt to infer the level distribution of users in real-world prediction interfaces like driving direction apps or advertising dashboards, so as to complement algorithmic solutions to problems like distribution shift with greater knowledge of how outcomes can arise. Future work might also pursue the use of synchronous experiments using platforms such as empirica.ly [1] where it is possible for post-decision feedback to combine decisions from all participants in real-time (e.g., [21]) to avoid the limitations of our staged design.

Our experiment also did not directly measure display reliance, but an interesting pursuit for future work would be to design a similar experiment where an agent's interest in continuing to use an information display is an outcome variable. For example, an experiment might elicit and quantify their willingness to pay for a display relative to the utility-optimal amount.

6 CONCLUSIONS

The distribution shift that occurs when presenting predictions poses challenges in designing an interface that can support data-driven decision-making in strategic settings. We presented the results of several large pre-registered online experiments to study the impact of design features like visualizing prediction uncertainty and prediction error on individual-level decision-making and aggregate-level system outcomes in a repeated congestion game, where agents' payoffs depended on their own actions and those of other agents viewing the same display. By endowing agents with varying level-k depths of thinking from a Poisson-CH model [8] in a novel staged experimental design, our work demonstrated how the design of a shared information display can affect agents differently depending on their level of strategic sophistication. The interface that provides post hoc decision feedback visualizing realized prediction error can help more sophisticated L2s to best respond and anticipate other agents' decisions. Our work also underlines the inherent trade-off that can arise between individual-level utility and collective outcome in strategic settings like congestion games. In the scenario we study, design choices that promote individual-level decision-making lead to worse collective outcomes in terms of lower percentages of the possible social welfare for the system being achieved, but, at the same time, improve users' trust and reliance on the prediction interface by decreasing distribution shift, making predictions more accurate post hoc. Our work opens up a new space of questions about what makes for a robust design strategy in settings where shared information displays may be inaccurate retrospectively.

REFERENCES

[1] Abdullah Almaatouq, Joshua Becker, James P Houghton, Nicolas Paton, Duncan J Watts, and Mark E Whiting. 2021. Empirica: a virtual lab for high-throughput macro-level experiments. *Behavior Research Methods* 53, 5 (2021), 2158–2171.

[2] Luc Anselin. 1988. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical analysis* 20, 1 (1988), 1–17.

[3] Luc Anselin. 1988. *Spatial econometrics: methods and models*. Vol. 4. Springer Science & Business Media.

[4] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68, 3 (2013), 255–278.

[5] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods* 10, 4 (2005), 389.

[6] Dirk Bergemann and Stephen Morris. 2019. Information design: A unified perspective. *Journal of Economic Literature* 57, 1 (2019), 44–95.

[7] Christopher J Bryan, Elizabeth Tipton, and David S Yeager. 2021. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour* 5, 8 (2021), 980–989.

[8] Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. 2004. A cognitive hierarchy model of games. *The Quarterly Journal of Economics* 119, 3 (2004), 861–898.

[9] Kathryn Chaloner and George T Duncan. 1987. Some properties of the Dirichlet-multinomial distribution and its use in prior elicitation. *Communications in Statistics-Theory and Methods* 16, 2 (1987), 511–523.

[10] Han-wen Chang, Yu-chin Tai, and Jane Yung-jen Hsu. 2010. Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining* 5, 1 (2010), 3–18.

[11] Giorgio Coricelli and Rosemarie Nagel. 2009. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences* 106, 23 (2009), 9163–9168.

[12] Diego Correa, Kun Xie, and Kaan Ozbay. 2017. Exploring the taxi and Uber demand in New York City: An empirical analysis and spatial modeling. In *96th Annual Meeting of the Transportation Research Board, Washington, DC*.

[13] Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2142–2151.

[14] Geoff Cumming and Sue Finch. 2005. Inference by eye: confidence intervals and how to read pictures of data. *American psychologist* 60, 2 (2005), 170.

[15] Sanmay Das, Emir Kamenica, and Renee Mirka. 2017. Reducing congestion through information design. In *2017 55th annual allerton conference on communication, control, and computing (allerton)*. IEEE, 1279–1284.

[16] Neema Davis, Gaurav Raina, and Krishna Jagannathan. 2016. A multi-level clustering approach for forecasting taxi travel demand. In *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)*. IEEE, 223–228.

[17] Xiaonan Dong, Xuehuan Li, Zongcheng Liu, and Qiuni Li. 2018. Non-cooperative game of multi-agent countermeasure systems based on cognitive hierarchy theory. In *Journal of Physics: Conference Series*, Vol. 1069. IOP Publishing, 012015.

[18] Peter Eades. 1984. A heuristic for graph drawing. *Congressus numerantium* 42 (1984), 149–160.

[19] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[20] Filippos Fotiadis and Kyriakos G Vamvoudakis. 2021. Recursive reasoning for bounded rationality in multi-agent non-equilibrium play learning systems. In *2021 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 741–746.

[21] Seth Frey and Robert L Goldstone. 2013. Cyclic game dynamics driven by iterated reasoning. *PloS one* 8, 2 (2013), e56416.

[22] William W Gaver. 1991. Technology affordances. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 79–84.

[23] Andrew Gelman, Jessica Hullman, and Lauren Kennedy. 2023. Causal quartets: Different ways to attain the same average treatment effect. *arXiv preprint arXiv:2302.12878* (2023).

[24] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. Bayesian workflow. *arXiv preprint arXiv:2011.01808* (2020).

[25] Michael F Goodchild. 1992. Geographical information science. *International journal of geographical information systems* 6, 1 (1992), 31–45.

[26] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 111–122.

[27] Rink Hoekstra, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review* 21 (2014), 1157–1164.

[28] Jake M Hofman, Daniel G Goldstein, and Jessica Hullman. 2020. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–12.

[29] Xiaohua Hu, Ronald Shonkwiler, and Marcus C Spruill. 2009. Random restarts in global optimization. (2009).

[30] Jessica Hullman. 2019. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 130–139.

[31] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2018. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 903–913.

[32] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.

[33] Shan Jiang, Le Chen, Alan Mislove, and Christo Wilson. 2018. On ridesharing competition and accessibility: Evidence from uber, lyft, and taxi. In *Proceedings of the 2018 World Wide Web Conference*. 863–872.

[34] Bryan D Jones. 1999. Bounded rationality. *Annual review of political science* 2, 1 (1999), 297–321.

[35] Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 99–127.

[36] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 892–902.

[37] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 5092–5103.

[38] Paula Kayongo, Glenn Sun, Jason Hartline, and Jessica Hullman. 2021. Visualization equilibrium. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 465–474.

[39] Pierre Legendre and Louis Legendre. 2012. *Numerical ecology*. Elsevier.

[40] Paul M Leonardi, Marleen Huysman, and Charles Steinfield. 2013. Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations. *Journal of computer-mediated communication* 19, 1 (2013), 1–19.

[41] Siyu Liao, Liutong Zhou, Xuan Di, Bo Yuan, and Jinjun Xiong. 2018. Large-scale short-term urban taxi demand forecasting using deep learning. In *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 428–433.

[42] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[43] Dean Lusher, Johan Koskinen, and Garry Robins. 2013. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

[44] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

[45] Richard D McKelvey and Thomas R Palfrey. 1995. Quantal response equilibria for normal form games. *Games and economic behavior* 10, 1 (1995), 6–38.

[46] Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. 2020. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems* 33 (2020), 4929–4939.

[47] John P Miller, Juan C Perdomo, and Tijana Zrnic. 2021. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*. PMLR, 7710–7720.

[48] Luis Moreira-Matias, João Gama, Michel Ferreira, and Luís Damas. 2012. A predictive model for the passenger demand on a taxi network. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 1014–1019.

[49] Trisalyn A Nelson and Barry Boots. 2008. Detecting spatial hot spots in landscape ecology. *Ecography* 31, 5 (2008), 556–566.

[50] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. Uncertain judgements: eliciting experts' probabilities. (2006).

[51] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

[52] José Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, Siem Heisterkamp, Bert Van Willigen, and R Maintainer. 2017. Package 'nlme'. *Linear and nonlinear mixed effects models, version* 3, 1 (2017).

[53] Brian W Rogers, Thomas R Palfrey, and Colin F Camerer. 2009. Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory* 144, 4 (2009), 1440–1467.

[54] Robert W Rosenthal. 1973. A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory* 2, 1 (1973), 65–67.

[55] Tim Roughgarden. 2005. *Selfish routing and the price of anarchy*. MIT press.

[56] Yossi Rubner, Leonidas J Guibas, and Carlo Tomasi. 1997. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA image understanding workshop*, Vol. 661. 668.

[57] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 1998. A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 59–66.

[58] Reinhard Selten. 1990. Bounded rationality. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft* 146, 4 (1990), 649–658.

[59] Herbert A Simon. 1956. Rational choice and the structure of the environment. *Psychological review* 63, 2 (1956), 129.

[60] Herbert A Simon. 1990. Bounded rationality. *Utility and probability* (1990), 15–18.

[61] Lisa M Sullivan, Kimberly A Dukes, and Elena Losina. 1999. An introduction to hierarchical linear modelling. *Statistics in medicine* 18, 7 (1999), 855–888.

[62] Ying Wen, Yaodong Yang, Rui Luo, and Jun Wang. 2019. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. *arXiv preprint arXiv:1901.09216* (2019).

[63] Ying Wen, Yaodong Yang, and Jun Wang. 2021. Modelling bounded rationality in multi-agent interactions by generalized recursive reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 414–421.

[64] GN Wilkinson and CE Rogers. 1973. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 22, 3 (1973), 392–399.

[65] James R Wright and Kevin Leyton-Brown. 2012. Behavioral game theoretic models: a Bayesian framework for parameter analysis.. In *AAMAS*, Vol. 12. 921–930.
[66] James R Wright and Kevin Leyton-Brown. 2017. Predicting human behavior in unrepeated, simultaneous-move games. *Games and Economic Behavior* 106 (2017), 16–37.
[67] Yinyu Ye. 1988. *Interior algorithms for linear, quadratic, and linearly constrained convex programming*. Stanford University.
[68] Dongping Zhang, Eytan Adar, and Jessica Hullman. 2021. Visualizing Uncertainty in Probabilistic Graphs with Network Hypothetical Outcome Plots (NetHOPs). *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 443–453.
[69] Kai Zhang, Zhiyong Feng, Shizhan Chen, Keman Huang, and Guiling Wang. 2016. A framework for passengers demand prediction and recommendation. In *2016 IEEE International Conference on Services Computing (SCC)*. IEEE, 340–347.

A FULL DESCRIPTION OF DATA PRE-PROCESSING AND COUNTERFACTUAL MODEL

We build our congestion game using the Chicago Taxi Trips data[4], which takes the form of origin-destination (OD) flows. Each observation represents a completed trip including variables such as taxi ID, start/end timestamps, and pickup/drop-off Community Areas (CAs)[5]. We leverage the findings from previous work on modeling taxi trips (e.g., [10, 16, 41, 48, 69]) and concepts from spatial econometrics [3, 25, 39, 49] to process the taxi trips data and formulate a counterfactual model that can support our strategic setting by computing individual player payoffs as well as the system outcome.

A.1 Defining the action set available to players

We first define the *action set* of our congestion game that each player must choose from by inspecting the average daily pickups by CAs between January 2014 and December 2015. We choose this time range because Chicago taxi demand was at its peak during this time: like many other major US cities, taxi pickups in Chicago had been declining annually because of the market competition from ride-sharing companies [12, 33].

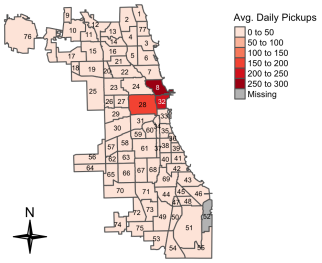

Fig. 10. Averaged daily pickups by Chicago CA.

As shown in Figure 10, the averaged daily pickups indicate spatial heterogeneity [2] as the spatial distribution of pickups is highly uneven. Three CAs in the Central Business District have significantly greater daily pickups than the rest of the CAs, and they are North Loop (CA = 8), West Loop (CA = 28), and the Loop (CA = 32). At the same time, the three contiguous spatial hot spots [49] show evidence of spatial dependency [25] in that the averaged daily pickups are similar in quantity. To control spatial heterogeneity and the number of total actions available in our strategic setting, we define a player's *action set* as these three spatial hot spots and refer to them as *action CAs*.

[4]The data is publicly available online at the Chicago Data Portal.
[5]Community Areas are a set of consistent boundaries that do not change over time and are also used for census collection. See detailed descriptions provided by Chicago's Department of Asset, Information and Services.

A.2 Reducing heterogeneity in training data

Taxi pickups can vary considerably depending on the time of day, day of the week, weather conditions, and special events (e.g,. sports games, holidays). While real taxi drivers could adjust their use of an information display in light of such factors, in our experimental setting, training a model without accounting for these irregular demand shocks could lead to highly variant display errors. We therefore implement a stratification scheme that groups taxi trips by start timestamps along two time dimensions: (1) days of a week (i.e., weekdays and weekends) and (2) time intervals of a day[6]. We define the pickup session of our strategic setting to take place at 9 AM, and select the strata containing 5.8 million trips from 522 weekday AM Peaks (7AM - 10AM) as relevant trips to infer each driver's 9 AM search decisions. We also account for two external factors that can influence taxi flows within the strata: the weather conditions and occurrence of special events. For weather conditions, we rely on four hourly meteorological variables[7] (i.e., apparent temperature, precipitation, snow depth, and wind speed). For special events, we consulted a list of Chicago CBD events[8] as well as national holidays. After conditioning on time period and external factors, we identified 221 (42%) homogeneous target weekdays from the trip strata.

A.3 Deducing search flows

Our objective is to have a counterfactual model to have a functional form pickups = f(flow), so that it can predict the total number of pickups (i.e., dependent variable) in each action CA at 9 AM, given a discrete flow distribution of drivers going to search over the three CAs (i.e., independent variable). Although we can aggregate the dependent variable by directly counting the total number of 9 AM pickups for each action CA from the taxi data, we face a hard data limitation that threatens the viability of our counterfactual model: the data only provides partial information about successful pickups or the *total supply* – we do not know the number of drivers that searched an action CA and did not get a pickup.[9] We therefore design an algorithm that can deduce drivers' search decisions using their prior pickup history during AM Peaks.

At a high level, the algorithm first identifies candidate drivers who might choose to search in each action CA at 9 AM on a target weekday by considering what it means *to be able to make a* pickup in an action CA at 9 AM on that weekday. By (1) back-tracing the previous drop-off CAs of drivers who found 9 AM pickup in an action CA and (2) computing the amount of idle time between their previous drop-off and 9 AM pickup, we can approximate the maximum amount of time drivers who obtained a pickup spent searching from their previous drop-off CAs. The output of this procedure is a collection of weighted edges, which we call ***trace dyads*** because they track drivers' movements on the target weekday in the form of $D \xrightarrow{T} P$ where D is the previous drop-off CA, P is an action CA that a driver found pickup in at 9 AM, and T is the weight representing the amount of idle time between previous drop-off and 9 AM confirmed pickup. For each unique trace dyad by D and P, we find the maximum idle time, t_{max}, and set t_{max} to be the time threshold used to identify candidate drivers who might choose to search in the action CA from each unique D. For example, if the data shows the maximum idle time for drivers who previously dropped-off in the North Loop before finding a 9:00 AM pickup in the Loop is 15 minutes on the target weekday, we include all unique drivers who dropped-off in the North Loop 15 minutes before 9 AM into our candidate list for that target weekday.

[6]We consulted the time intervals defined by Uber Movement, which are AM Peak (7AM-10 AM), Midday (10AM-4PM), PM Peak (4PM-7PM), Evening (7PM-12AM), and Early Morning (12AM-7AM).
[7]Hourly meteorological variables are collected from Weatherstack API.
[8]Data is publicly available by Chicago's Department of Cultural Affairs and Special Events.
[9]We also do not observe total demand, but the convention is to use total pickups from a location as a proxy of total demand.

The algorithm next classifies each candidate driver identified by trace dyads into three types according to their 9 AM vacancy status shown by the taxi data. These are (1) those who successfully found a pickup in an action CA at 9 AM, (2) those who successfully found pickups elsewhere at 9 AM, and (3) those who failed to find a pickup and have no trip history in the data at 9 AM. Drivers of type (1) must have searched in one of the action CAs and hence are included in the search flow to the action CA where they found a pickup. Drivers of type (2) are removed from the candidate list. Drivers of type (3) are of primary interest: we need to deduce where they might have searched and decide if they should be included in the search flow to any action CA.

The algorithm consults each type (3) driver's search prior and tabulates unique consecutive pickups completed by the driver in the AM Peaks of the past tens days prior to the target weekday. From each driver's search prior, we create a collection of ***search dyads*** expressed as $V \xrightarrow{N} S$ where V is the dropoff CA of the previous trip, S is the pickup CA of the consecutive trip, and the weight N is the number of occurrences of the pickup pattern. A search dyad can be interpreted as: in the AM Peaks of the past 10 days, given a driver dropped off in v, she found her next pickup in s for n times. Since the collection of search dyads with the same V can be seen as proxies of a driver's conditional search preference from V, the algorithm identifies the S with the maximum N as the search CA for the driver.

A.4 Counterfactual Model

The above pre-processing steps result in a training dataset that contains 9 AM flow-pickup pairs on 221 homogeneous weekdays for each candidate CA, summarized from 2.1 million relevant trips completed by 5109 Chicago taxi drivers. Each observation from the processed data describes the number of observed pickups in an action CA, resulting from the quantity of deduced flow going into the action CA on each of the 221 weekday. We use a Bayesian Multilevel model to estimate pickup distributions and specify the model in Wilkinson-Rogers-Pinheiro-Bates syntax [52, 64] as follow:

1: $lpickup \sim \text{Gaussian}(\mu, \sigma)$
2: $\mu = lflow + (1 + lflow \mid CA)$

As shown, *lpickup* is the dependent variable and *lflow* is the independent variable, both are log-transformed before model fitting. *CA* is a categorical variable used as the identifier for action CAs. In line 1, we define the likelihood function of *lpickup* to follow a Gaussian distribution, which is parametrized by μ (mean) and σ (scale). In line 2, we specify a hierarchical model with varying intercepts and slopes. This is because although pre-processing (described in Appendix A.2) can reduce heterogeneity of pickup-flow dynamics within action CAs, each action CA can still have intrinsically different dynamics converting flows to pickups. The intercept and slopes can co-vary because if an action CA has larger average pickups (i.e., high intercept), it could signal relatively stronger rate of pickups conversion from flows. Because the action CAs we defined are neighbors to each other subject to the effect of spatial dependence, the dynamics learned by the model from an action CA can be used to improve estimate about the other CAs. Therefore, we want to model the covariance in flow-pickup dynamics as a result of spatial dependence between the action CAs, and improve CA-specific estimates through information pooling [44, 61]. Detailed model diagnostics and posterior predictive checks can be found in the Supplemental Material.

A challenge in creating a counterfactual model is that the simulated and observed flows we pass to the model to generate the display and to score decisions may in some cases fall below or exceed the flow range observed from the historical data. We use two heuristics to ensure the model makes reasonable predictions in such cases:

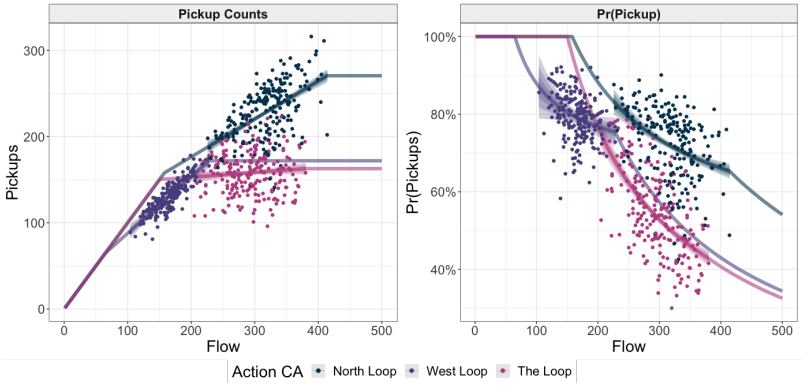

Fig. 11. Counterfactual Model Fits

- If a flow to an action CA is greater than the maximum historical flow observed from the data, we use the model to compute an action CA's expected pickups using the historical maximum flows. This captures how predicted pickups would at some point stop increasing with greater flow.
- If a flow is less than the minimum historical flow, we use the counterfactual model to predict pickups. Once predicted pickups exceeds the simulated flow, we assign a 100% pickup probability.

We present our model fits in Figure 11 with the dependent variable in counts (as defined in the model) on the left and converted pickup probability on the right. When more drivers choose to search in an action CA, the pickup probability becomes lower on average, though with varying slopes depending on the CA. The predictions become deterministic only when flows fall outside of the range observed historically from the data.

B DECISION SCENARIOS

Table 6. The orders of weekdays (decision scenarios) used to create our repeated game for the main experiment and robust composition check, and the robust trial order check, with the corresponding best responses by trial, level, and treatment.

Trial Order		Date	Level-1	Level-2			
Main+Robust composition	Robust trial order			Static+Bandit	Static+Full	NetHOPs+Bandit	NetHOPs+Full
0	0	2014-07-22	West	East	East	East	East
1	6	2014-05-13	West	East	East	East	East
2	14	2015-05-21	West	East	East	East	East
3	10	2014-06-05	West	East	East	East	East
4	7	2015-06-25	North	East	East	East	East
5	12	2014-11-07	West	East	East	East	East
6	15	2015-06-30	West	East	East	East	East
7	1	2015-10-26	West	East	East	East	East
8	9	2015-10-22	West	East	East	East	East
9	4	2015-10-14	North	East	East	East	East
10	3	2014-09-26	West	East	East	East	East
11	8	2014-12-05	West	East, North	East, North	East	East
12	5	2015-08-28	West	East	East	East	East
13	13	2014-05-22	West, North	East	East	East	East
14	2	2014-10-24	West	East	East	East	East
15	11	2014-09-29	West	East, North	East	East	East

C GLOSSARY

Table 7. Glossary of terms and concepts used in empirical game theory in the context of our multi-agent strategic setting.

Game Theory	
Congestion game	A broad class of non-cooperative games where each action represents a congestible good and is associated with a cost function, which incurs cost that increases with the number or fraction of agents who chose the same action.
Principal	A service or prediction provider. In our experimental setting, we assume the role of the principal, or the taxi company.
Agents	Users of predictions who also make decisions. In our experimental setting, participants act as taxi drivers who use the information display to inform their search decisions.
Action set	The collection of all possible actions available to a player in a game. In our experiment, action sets contain three districts, referred to as *action CAs*, that a participant can choose to search for passengers.
Payoff-relevant information	Pertinent details that influence potential rewards or outcomes. In our experiment, the payoffs are influenced by two main factors participants observe on the information display: (1) deduced flow, (2) predicted pickup probability.
Experimental Setting	
Poisson Cognitive Hierarchy Model	A behavioral model we utilize to characterize agents' strategic sophistications through levels, with proportions following a Poisson distribution.
Level-k framework	In a level-k framework, all agents are assumed to be myopic. They believe they are the most sophisticated agents in action. They assume all other competing agents are distributed according to a normalized Poisson distribution for levels ranging from 0 to $k-1$.
Level distribution	A Poisson distribution including L0-L2 drivers representing the "true" population mixture over levels, which is exclusive knowledge of the principal used to aggregate the system outcomes by combining decisions using the taxi data and the collected responses from our participants.
Level endowment	Endowing levels by informing participants of the level mixtures. After normalizing the level distribution, this helps them perceive the rest of the population as consisting entirely of agents of levels lower than their own.
Realized prediction error	The discrepancy between the predicted outcome and the realized outcome.
Level-specific outcome	An outcome that aligns with participants' endowed levels to score their decisions and provide feedback in each trial.
System outcome	An outcome derived by integrating decisions from all levels, in accordance with the level distribution that defines the population.
Experimental Manipulations	
Static display	An information display that presents predictions as a static point estimate.

Table 7 continued from previous page

NetHOPs	A frequency-based uncertainty visualization technique that displays predictions and communicates the associated uncertainty via animated frames.
Bandit feedback	A feedback mechanism that solely informs participants about whether they secured a pickup based on their decision.
Full feedback	A comprehensive feedback display that not only informs participants of the decision outcome but also visualizes the realized prediction error.
Response Variables	
Best response	A response variable assessing whether a participant chose the district yielding the highest expected pickup probability, as dictated by their endowed level.
Anticipation error	A response variable evaluating a participant's ability to foresee other participants' search decisions, determined using the Earth Mover's Distance.
Distribution shift	A response variable that quantifies the disparity between the deduced search flows displayed to participants and the flows derived from system outcomes.
Social welfare ratio	A response variable that computes the proportion of social welfare achieved (i.e., total realized pickups) out of the total possible social welfare obtainable (i.e., maximum pickups).

Received 4 October 2023