User Strategization and Trustworthy Algorithms

Sarah H. Cen, Andrew Ilyas, and Aleksander Mądry

Abstract

Many human-facing algorithms—such as those powering recommender systems or hiring decision aids—are trained on data provided by their users. A common assumption of these algorithms is that user behavior is *exogenous*: that is, a user would react to a given prompt (e.g., a recommendation or hiring suggestion) in the same way no matter what algorithm generated it. For example, latent factor models in recommender systems posit that a user's interest in an item depends on the features of the user and of the item, but not of the model itself. In practice, however, user behavior is not exogenous—users are *strategic*. Recent studies document, for example, TikTok users changing their scrolling behavior after learning that TikTok uses it to curate their feed, and Uber drivers changing how they accept and cancel rides in response to changes in Uber's algorithm.

Our work studies the implications of strategization by modeling the interactions between a user and their data-driven platform as a repeated, two-player game. We leverage results from misspecified learning to characterize the effect of strategization on data-driven algorithms. We show that although strategization can actually help platforms in the short term, it ultimately corrupts platforms' data and hurts their ability to make counterfactual decisions. We connect this phenomenon to user trust, and show that designing trustworthy algorithms can go hand in hand with accurate estimation. We provide a formalization of trustworthiness that inspires potential interventions.

1 Introduction

In the age of personalization, data-driven platforms have become increasingly reliant on data provided by their users. Platforms like Facebook, Amazon, and Uber tailor their services to each user based on the user's interaction history. Even data-driven tools in medicine and hiring utilize previous interactions in order to fine-tune and improve their results.

Traditionally, platforms make a key assumption when processing user data: that user behavior is *exogeneous*, i.e., how a user behaves depends on the user but *not* on the platform's algorithm. In practice, however, users are not blind to how their platforms operate. Rather, users often adapt their behavior—or *strategize*—based on their perception of how the platform works. For instance, a Facebook user might not click on a post not becauase they find it uninteresting, but because they believe the algorithm will over-recommend similar content in the future if they do click. Or an Uber driver may cancel low-paying rides that they would normally accept because they have learned that Uber's algorithm does not penalize them for excessive cancelations [Mar20].

While it is convenient to assume that user behavior is exogeneous, it can mislead the platform. When users strategize, it becomes difficult to discern whether a user action (e.g., skipping a video on YouTube) is exogenous (the user would skip this video, regardless of YouTube's algorithm) or *endogenous* (the user skips because they believe watching this video will trigger undesirable recommendations under YouTube's algorithm). As a result, the platform's data is no longer generalizable—data gathered under one algorithm cannot be used to make reliable predictions about outcomes under a different algorithm.

Aware that users adapt their behavior to their platform's algorithm, some platforms have begun taking actions that lead to a cat-and-mouse game between the platform and user. For instance, Facebook may notice that users strategize by selectively clicking the "like" button and begin tracking how quickly users scroll down their feeds, using the dwell time on each post as proxies for user interest. Upon learning that Facebook is tracking their dwell time, users may begin strategizing how they scroll, prompting Facebook to respond yet again and resulting in a cycle that only serves to erode the trust between a user and their platform.

It is natural to ask: is this outcome inevitable? Is out-maneuvering users the only way for platforms to obtain high-quality data? Not necessarily. It turns out that trustworthy algorithm design plays a key role. When an algorithm is trustworthy, users do not feel compelled to strategize—they provide data that, for all intents and purposes, *looks exogenous*. We formalize this intuition, calling a platform *κ-trustworthy* when their choice of algorithm does not induce strategization and ensures users receive at least κ utility on average. Intuitively, a platform is trustworthy when a user believes that the platform looks out for their interests along the same axis that would induce the user to strategize. We connect this formalization to the existing literature on trustworthiness, comparing it to definitions that arise in political science and law. Using this framework, we show that trustworthy design can mitigate the effects of strategization on a platform, to the benefit of both the platform and user.

1.1 Summary of contributions

We begin by modeling the interactions between a user and their data-driven platform as a repeated, two-player game (Section 3). Under our model, there are two agents: a user and their platform. At each time step, the platform puts forth a proposition (e.g., a prediction or a recommendation), the user responds with a behavior (e.g., agreeing with the prediction or ignoring the recommendation), and each party receives a payoff based on their action. The platform's goal is to generate high-payoff propositions by developing a good estimate of the user's behavioral tendencies from repeated interactions with the user. In our model, the platform "moves first" in that it declares how it generates propositions, after which the user "moves second" by choosing how they respond to propositions (akin to a Stackelberg game). Importantly, this model allows users to observe and adapt to the platform's algorithm.

We then define what it means for a user to be strategic in Section 4. We call a user "naive" if they adopt a best-response strategy, i.e., at every time step, they behave as though they are only interacting with the platform once. In contrast, we call a user "strategic" if they anticipate how their next action may affect the platform's future propositions and, in turn, the user's long-term payoff. That is, the strategic user understands that their actions can affect the platform's long-term behavior and adapt accordingly. In order to study the long-term behavior of the two-player system, we utilize the notion of a *globally stable set* [FII20]. Typically, characterizing how users plan ahead involves solving a difficult optimization over a discount reward function. However, using the globally stable set allows us to bypass this calculation by examining the system's behavior at equilibrium.

We show in Section 6 that user strategization can both help and hurt the platform. We first show that strategization can *improve* the platform's payoff (as long as the platform does not change its algorithm). Intuitively, if the user's and platform's payoffs are sufficiently aligned, then strategization can help overcome deficiencies in the algorithm. We then show that, although strategization can benefit the platform when its algorithm is fixed, it can *distort the data that the platform* *collects and mislead the platform*. We first demonstrate that, if the platform collects data when a user is strategic, its estimate of how it will perform under a new algorithm can be arbitrarily poor. We further show that, when a user is strategic, expanding the hypothesis class (or model family) that the platform uses to estimate user behavior can lower the platform's payoff, even if everything else is held fixed. This finding is counterintuitive as expanding one's hypothesis class typically results in better estimation. Although our negative results focus on the platform's ability to estimate its payoff, they extend to other estimation tasks—the takeaway is that strategization can distort the data that the platform may use for a variety of tasks. We then show that *these difficulties disappear when the user is not strategic*, i.e., when the user does not plan ahead, which suggests that platforms should design algorithms that do not incentivize users to strategize.

In the face of user strategization, platforms are left with a few options. They could simply use the collected data and risk drawing incorrect inferences. They could post-process the data, but "correcting" the data by removing the effect of strategization is extremely challenging in complex settings because the user can strategize for many different reasons and along many different axes. Alternatively, platforms could "correct" for strategic behavior by gathering more data. In Section 7, we discuss when and how these approaches fall short.

We then discuss how another approach—designing trustworthy algorithms—can improve user and platform outcomes. Formally, we define a κ-trustworthy algorithm as one that (i) does not incentivize users to strategize and (ii) guarantees that the user's payoff is at least $\kappa \geqslant 0$. We connect this definition to existing sociological notions of trust, e.g., trust as "encapsulated interest" [Har06]. We conclude by enumerating three reasons users strategize and two interventions for trustworthy design: offering multiple algorithms and providing feedback mechanisms.

2 Related work

Our work draws inspiration from extensive lines of work spanning computer science, economics, game theory, and the social sciences. We outline a few of the most related areas to our work below.

Endogenous learning. One of the key aspects of our model is *endogeneity* on the side of the platform, i.e., the platform's actions affect the data it collects. There is a vast literature in both economics and computer science that studies endogenous learning, some of which we heavily rely on in this work.

Our work draws from a long line of work at the intersection of economics and game theory that explores the dynamics of *endogenous misspecified learning*. Although we do not explicitly rely on it here, the Berk-Nash equilibrium concept [EP16] provides a basis for much of this work. The concept was later refined, expanded, and improved upon in several directions [FRS17; FLS21a; Boh16; BH21; FII20]. In this work, we rely most heavily on the results of Frick et al. [FII20] due to their generality, but our results are portable in the sense that new ways of characterizing globally stable beliefs will allow for even sharper results in our setting.

There are also many related equilibrium concepts in the Economics literature that we do not explore in this work, such as self-confirming equlibrium [FL93] and subjective equilbrium [KL93; KL95]. Relating these models (and others) to our setting is an interesting question for future work.

Several works in the computer science literature also study misspecified learning. Of these, the most related is the work of Perdomo et al. [Per+20] which introduces the idea of *performative prediction*. The performative prediction setup mirrors that of supervised learning, except that the learner's current parameter estimate θ affects the data distribution D_θ. Our model can be viewed as an instance of performative prediction (which in turn, can be viewed as an instance of reinforcement learning [BHK22]) that focuses on a specific kind of performativity induced by users adapting to their platforms.

Strategic classification and Stackelberg games. Strategic classification [Har+16] is a two-player game in which one player (the *decision-maker*) deploys a machine learning algorithm, and the other player (the *decision subject*) strategically reports their features to attain a favorable decision. Strategic classification also features endogenous learning, and is a special case of the performative prediction setup mentioned above. Strategic classification has been the subject of a deluge of recent work in computer science [BKS12; Har+16; Don+18; Gha+21; Zrn+21]. Broadly, our model differs from one of strategic classification in that (a) we restrict the platform to a specific learning algorithm; (b) users have their own utility functions that they can optimize arbitrarily, and are not bound to small perturbations of some "ground-truth" features; (c) users can be myopic or non-myopic in how they interact with the platform.

Prior works have studied deviations from these assumptions. For (a), Zrnic et al. [Zrn+21] study the case where rather than reacting instantaneously, the decision subject is *learning* at the same time as the the decision-maker. Closer to our work, Levanon and Rosenfeld [LR22] introduce *generalized* strategic classification, where the decision subject has a utility function that can be more aligned with that of the platform. Finally, Haghtalab et al. [Hag+22] study Stackelberg games (which generalize strategic classification) with *non-myopic agents* (i.e., where agents seek to optimimize their long-term, rather than immediate, payoff). As previously discussed, our model differs from theirs in that in our case users optimize their expected *limiting* payoff, which, and so delaying user input has no bearing on the game dynamics.

User strategization on recommender systems. There are several related lines of work to ours that concern user strategization on recommender systems specifically. One line of work studies the ways in which users try to influence (and succeed in influencing) their recommendations on popular platforms [SVM22; KL23; SHS22]. Closer to our work are theoretical models of user strategization on recommender platforms [CIM22; HHP23], where [CIM22] is an earlier version of this work. Compared to the latter of these works, we propose a model that (a) extends more generally to data-driven platforms; (b) allows us to study the causes and effects of strategization on the platform and its ability to make counterfactual inferences; and (c) connects strategization back to trustworthiness of the platform.

Another related work is that of Kleinberg et al. [KMR22], who study a model of *inconsistent preferences* under which users have a "myopic" system 1 that consumes addictive content and a "non-myopic" system 2 that considers the value of content. Although we also study myopic and non-myopic users, we use these terms in way that is subtly different from [KMR22], and their focus (inconsistent preferences) is neither necessary nor sufficient for strategization to emerge. Additionally, strategization concerns users' response to their platforms' algorithm, whereas in [KMR22] the platform's algorithm is fixed.

Multi-agent learning and games. There are also many lines of work on multi-agent learning, including multi-agent reinforcement learning (see [ZYB21; Tan93; BBD08] and their references), multi-agent Bayesian learning (e.g., [WAO20; WAO21]), and inverse reinforcement learning (particularly its cooperative [Had+16] and adversarial [YSE19] variants). All of these models capture agents learning and interacting with each other at varying levels of generality. Here, we focus on a particular instantiation of such setups where two agents interact in a very specific way, mirroring the interaction between a user and a data-driven platform. Our interests are also more specifically in studying the causes and effects of strategic behavior in this setup.

Another related work studies repeated alternating two-player games [Rot+10] and considers the complexity of computing equilibria in such games. In our work, we avoid having to compute such equilibria by assuming that the platform follows a fixed Bayesian updating strategy.

Mechanism design and incentive-compatible auctions. Our notion of "trustworthiness" from Section 7.1 is highly related to the idea of incentive-compatibility in mechanism design. There is a rich line of work in econometrics on designing incentive-compatible methods for repeated games in general (see, e.g., [MS06; BM93; KL93]), and more specifically for repeated auctions [Den+21; Abe+19; Ned+22; ARS13; KN19]. In the latter, a long line of work stems from the tight connections between incentive-compatibility and differential privacy [MT07; NST12].

Human-AI collaboration. Finally, our work contributes to a growing body of work on Human-AI collaboration [Wan+19; Wan+20; Moz+22] and teaming [Zha+21; Ban+19], which both study primarily collaborative interactions between AI systems and their users. In particular, one can view our observations about conditions for strategization (and corresponding recommendations about trustworthy algorithm design) in Section 7.2 as principles for human-AI collaboration in the case where the AI and the human are not entirely aligned.

Within Human-AI collaboration, one line of work in particular isolates *trust* as a building block of successful interaction (see, e.g., [OY20; HHC23; Eze+19; Bao+21] and references therein). Our work supports the high-level idea that designing systems with trust in mind is important, but we explore a slightly different notion of trust than what is typically considered. In particular, in our case trustworthiness dictates whether a user will maximize their immediate utility at each step of a game. By comparison, human-AI collaboration usually explores trust to the end of getting users to use AI systems in the first place (or comfortably delegate complext tasks to AI systems).

3 Model

In this section, we present our game-theoretic model of the interactions between a user and their data-driven platform. At its core, the model comprises a repeated, two-player game. At every time step, the platform generates *propositions* (e.g., recommendations). The user then responds to these propositions (e.g., chooses whether to engage with a recommendation) with *behaviors*.

A key feature of our model is that there is no "ground-truth" user behavior. In particular, the way that the user responds to propositions may depend on *how* the platform generates them. Our model therefore departs from earlier works that assume a user's behaviors are drawn from a single fixed, unknown distribution.

Since in this work we are interested in how users adapt to their platforms, we study the setting where the platform first declares the strategy it will use to generate propositions. The user then decides on how they wish to behave, with full knowledge of the platform's intended strategy. [1]

3.1 Setup

We model the interactions between a user and their platform as a two-player game. Formally, for some $d_1, d_2 > 0$, let $\mathcal{Z} \subset \mathbb{R}^{d_1}$ and $\mathcal{B} \subset \mathbb{R}^{d_2}$ denote the action spaces of the platform and user, respectively, where we assume that $\mathcal{B}$ is finite. Then, at every time step $t = 0, 1, 2, \ldots,$

[1] While full user knowledge is a strong assumption, the effects we discuss here will only be exacerbated when users have an imperfect model of the platform. We leave studying this case to future work.

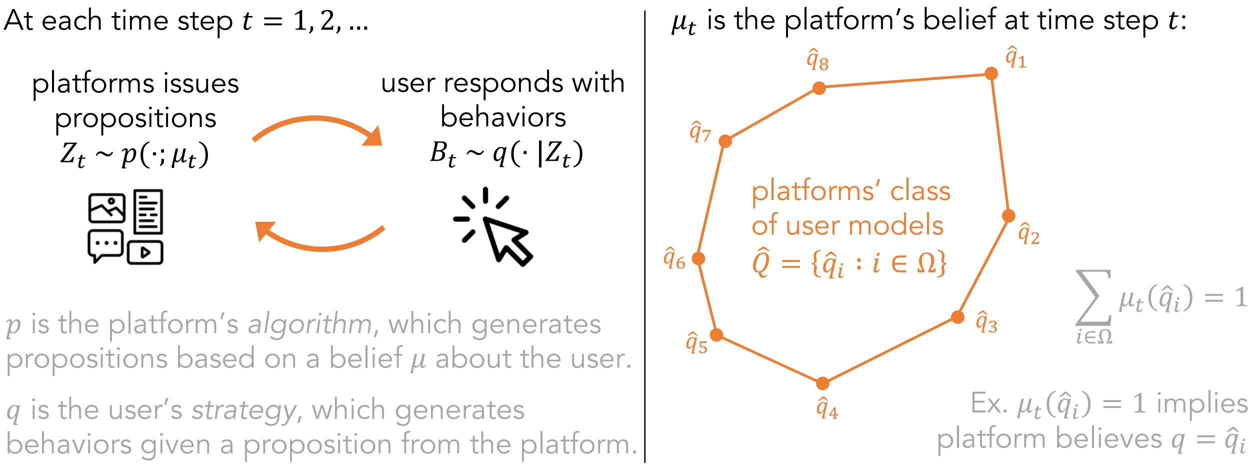

At each time step $t = 1, 2, \ldots$
platforms issues propositions $Z_t \sim p(\cdot; \mu_t)$
user responds with behaviors $B_t \sim q(\cdot \mid Z_t)$
p is the platform's *algorithm*, which generates propositions based on a belief μ about the user.
q is the user's *strategy*, which generates behaviors given a proposition from the platform.
μ_t is the platform's belief at time step t:
$\hat{q}_1$
$\hat{q}_2$
$\hat{q}_3$
$\hat{q}_4$
$\hat{q}_5$
$\hat{q}_6$
$\hat{q}_7$
$\hat{q}_8$
platforms' class of user models
$\hat{Q} = \{\hat{q}_i : i \in \Omega\}$
$\sum_{i \in \Omega} \mu_t(\hat{q}_i) = 1$
Ex. $\mu_t(\hat{q}_i) = 1$ implies platform believes $q = \hat{q}_i$

Figure 1: Illustration of the setup described in Section 3. (Left) At each time step t, the platform issues propositions Z_t, and the user responds with behaviors B_t. The user's actions are determined by their strategy $q : \mathcal{Z} \to \Delta(\mathcal{B})$, and the platform's are determined by the algorithm p and the hypothesis class $\widehat{\mathcal{Q}}$. (Right) The platform's actions at time t depend on its belief μ_t. Here, μ_t is a distribution (i.e., set of weights) over $\widehat{\mathcal{Q}}$ such that $\mu_t(\hat{q}_i)$ denotes the probability that the platform assigns to $q = \hat{q}_i$ at time t.

1. The platform generates *propositions* $Z_t \in \mathcal{Z}$.
2. The user responds with *behavior* $B_t \in \mathcal{B}$, drawn from the conditional distribution $q(\cdot|Z_t)$.
3. The platform and user collect payoffs $V(Z_t, B_t)$ and $U(Z_t, B_t)$, respectively.

This setup is given on the left side of Figure 1. We use the shorthand $q : \mathcal{Z} \to \Delta(\mathcal{B})$ to denote the set of conditional distributions $\{q(\cdot|Z) : Z \in \mathcal{Z}\}$. the user's *strategy*, i.e., a mapping from propositions Z to behavior distributions $q(\cdot|Z)$ for all $Z \in \mathcal{Z}$. Throughout this work, $\Delta(X)$ denotes the simplex over a probability space X. Furthermore, we assume that the payoffs V and U as well as the action spaces $\mathcal{Z}$ and $\mathcal{B}$ are *exogeneous*, i.e., they are pre-specified. We assume that the payoff functions U and V are bounded, and that everything is scaled such that $U, V \in [0, 1]$.

Generating propositions. To generate propositions, the platform first constructs an estimate of the user's strategy $q \in \mathcal{Q}$. In particular, the platform assumes that q belongs to some set $\widehat{\mathcal{Q}}$, which we assume to be finite. $\widehat{\mathcal{Q}}$ can be thought of as the platform's hypothesis class, with each $\hat{q}_i \in \widehat{\mathcal{Q}}$ being one of the platform's "user models." The platform constructs its estimate of q using a *belief* $\mu_t \in \Delta(\widehat{\mathcal{Q}})$ over the hypothesis class $\widehat{\mathcal{Q}}$. For instance, $\mu_t(\hat{q}_i) = 1$ means that the platform believes $q = \hat{q}_i$ with full certainty at time step t. When there is no user model that matches the user's chosen strategy (i.e., $q \notin \widehat{\mathcal{Q}}$), we say that the platform is *misspecified*.

At each time step, the platform uses a *(proposition) algorithm* $p : \Delta(\widehat{\mathcal{Q}}) \to \Delta(\mathcal{Z})$ to map its belief μ_t to a distribution over propositions. That is, at each time t, the platform uses its current belief μ_t to sample a proposition $Z_t \sim p(\cdot; \mu_t)$, as shown on the right side of Figure 1.[2] Intuitively, the algorithm p captures whether the platform chooses to maximize revenue, social welfare, or any other objective (based on its current belief μ_t).

[2] One can think of μ_t as parameters of a machine learning model trained on the data gathered until time step t.

Bayesian updating. We focus on the case where the platform uses Bayesian updating to maintain its belief μ_t. Specifically, the platform starts with a full-support initial belief $\mu_0 \in \Delta(\widehat{\mathcal{Q}})$, i.e., a distribution that assigns non-zero probability to every possible model $\hat{q} \in \widehat{\mathcal{Q}}$. At each time step t, the platform observes the user's behavior B_t in response to the proposition Z_t, then updates its belief μ_t to μ_{t+1} using Bayes' rule, i.e., the platform applies the update

$$\mu_{t+1}(\hat{q}_i) = \frac{\mu_t(\hat{q}_i) \cdot \hat{q}_i(B_t|Z_t)}{\sum_{j\in\Omega} \mu_t(\hat{q}_j) \cdot \hat{q}_j(B_t|Z_t)}, \qquad \forall\, \hat{q}_i \in \widehat{\mathcal{Q}}. \tag{1}$$

Committing to strategies. Thus far, we have instantiated a repeated, two-player game between a user and their platform. The platform generates propositions by applying its *algorithm* p to its (evolving) belief $\mu_t \in \Delta(\widehat{\mathcal{Q}})$ in order to sample $Z_t \sim p(\cdot; \mu_t)$. The user responds with behaviors $B_t \sim q(\cdot|Z_t)$. Note that, for a fixed action spaces, payoffs, and initial beliefs $(\mathcal{B}, \mathcal{Z}, U, V, \mu_0)$, the user's and platform's actions are fully determined by q, p, and $\widehat{\mathcal{Q}}$. We therefore refer to q as the *user's strategy* and the tuple $(p, \widehat{\mathcal{Q}})$ as the *platform's strategy* (see Table 1).

In this work, we are interested in how users adapt to their platforms. We therefore study the setting in which the platform commits to a strategy $(p, \widehat{\mathcal{Q}})$ at the start of the game, after which the user chooses their strategy q, which may depend on $(p, \widehat{\mathcal{Q}})$. In order to characterize user adaptation, we study the idealized setting in which the user has perfect knowledge of $(p, \widehat{\mathcal{Q}})$.

3.2 Examples

Our model captures a variety of data-driven settings. For instance, the interactions between a user and their recommender system (e.g., Netflix or Facebook) can be viewed as a repeated, two-player game. (Indeed, we provide a detailed recommender system example in Section 5). Our model also captures other contexts, such as data-driven hiring and ride-share matching, as detailed below.

Example 3.1 (Hiring example)**.** *In the hiring context, an employer (i.e., the user) uses a data-driven hiring platform to determine which job candidates to interview. At each time step t, Z_t denotes the set of candidates at time t and the scores that the hiring platform assigns to the candidates. B_t denotes the employer's decisions (i.e., which candidates are interviewed) and the final outcomes (i.e., who is hired). The employer's preferences over types of employees are therefore captured by $q(\cdot|\cdot)$. The hiring platform receives payoff $V(Z_t, B_t)$ based on the scored candidates Z_t and the employer's decisions B_t—for example, V might represent a pay-per-success scheme, scaling with how many top-ranked applicants are successfully hired. Meanwhile, U is the employer's payoff, e.g., how many people it successfully hires using the tool.*

The hiring platform's goal is to learn the employer's preferences so that it can provide scores attuned to the employer's needs. It does so by first assuming that the employer's decision mechanism q belongs to a hypothesis class $\widehat{Q}$, which describes the possible employers that the platform can model. As it receives more data on the employer's hiring practices, it updates its belief μ_t about the employer's preferences. Given its current belief μ_t, the platform presents candidates by sampling from a distribution $p(\cdot; \mu_t)$, where we call p the algorithm that the hiring platform uses. For instance, p might select the candidates it believes the company will be most likely to hire (according to μ_t) subject to equity constraints.

Example 3.2 (Uber example)**.** *In the ride-sharing context, drivers use a ride-sharing app—say, Uber—to find riders. At each time step, Uber proposes a ride Z_t and the driver decides (using their behavior B_t) whether to accept, reject, or accept then cancel the ride. Uber's payoff from the interaction is $V(Z_t, B_t)$, which could be a constant fraction of the ride's cost if the driver accepts it. Similarly, $U(Z_t, B_t)$ is the driver's payoff, which might be zero if they decline the ride, and might otherwise depend on whether the ride takes the driver closer to home, whether the ride involves unpleasant driving areas, among other factors.*

Uber's goal is to match drivers and riders. To minimize delays, Uber attempts to learn each driver's habits and preferences. Specifically, it assumes that the driver's decision function $q(B_t|Z_t)$ for whether to accept, decline, or cancel a ride belongs to some class $\widehat{\mathcal{Q}}$. After observing the driver's behavior, Uber updates its belief μ_t about q, and then uses its belief along with its algorithm p to better match riders and drivers.

4 User strategization

In our model, as given in Section 3, the user selects their strategy q after the platform has declared its strategy $(p, \widehat{\mathcal{Q}})$. Although our model allows us to analyze a wide range of user behaviors, there are two types of users of particular interest. The first type—a *naive* user—behaves as though they are only interacting with the platform once by playing actions that maximize their immediate payoff. On the other hand, a *strategic* user plans ahead; they adapt their strategy to the platform's strategy $(p, \widehat{\mathcal{Q}})$ in order to elicit high payoffs in the long run. In this way, a strategic user's behavior is dependent on $(p, \widehat{\mathcal{Q}})$, whereas a naive user's behavior remains the same across different choices of $(p, \widehat{\mathcal{Q}})$. We formalize these two types of users below and visualize their behavior in Figure 3.

4.1 Naive user

At each time step t, the naive user plays as though they are only interacting with the platform once by choosing the action B_t that maximizes their payoff U under the given proposition Z_t, as defined next. If multiple behaviors $B \in \mathcal{B}$ maximize the immediate payoff, we assume the user chooses between them uniformly at random.

Definition 4.1 (Naive user)**.** *The naive user adopts the strategy q^{BR}, defined as*

$$q^{BR}(B|Z) \propto \mathbf{1}\{B \in \arg\max_{B \in \mathcal{B}} U(Z, B)\}, \qquad \forall\, B \in \mathcal{B}, Z \in \mathcal{Z}.$$

Importantly, a naive user's strategy q^{BR} is independent of the platform's strategy; that is, q^{BR} remains the same across all choices of $(p, \widehat{\mathcal{Q}})$.

Table 1: Key concepts and notation from Sections 3 and 4.

Object	**Symbol**	**Description**
Proposition space	$\mathcal{Z}$	Platform action space, subset of $\mathbb{R}^{d_1}$ (exogeneous)
Behavior space	$\mathcal{B}$	User action space, finite subset of $\mathbb{R}^{d_2}$ (exogenous)
Payoff functions	U, V	Function that maps $\mathcal{Z} \times \mathcal{B}$ to $\mathbb{R}$ (exogeneous)
User strategy	q	Mapping $q : \mathcal{Z} \to \Delta(\mathcal{B})$ such that $B_t \sim q(\cdot|Z_t)$
Hypothesis class	$\widehat{\mathcal{Q}}$	Finite set of models $\{\hat{q}_i : i \in \Omega\}$ that platform uses to estimate q
Platform belief	μ_t	Distribution over $\widehat{\mathcal{Q}}$
Platform algorithm	p	Function that maps a belief $\mu \in \Delta(\widehat{\mathcal{Q}})$ to a distribution in $\Delta(\mathcal{Z})$ such that $Z_t \sim p(\cdot; \mu_t)$

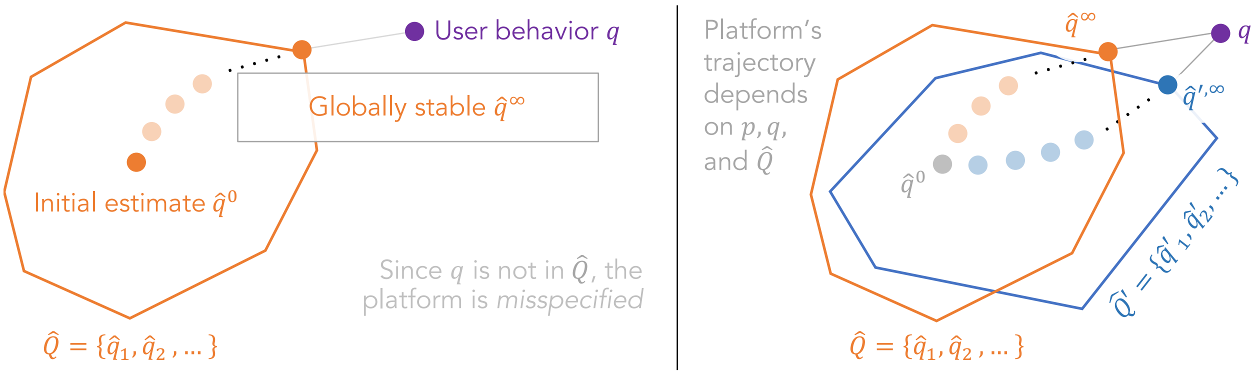

User behavior q
Globally stable $\hat{q}^\infty$
Initial estimate $\hat{q}^0$
Since q is not in $\hat{Q}$, the platform is *misspecified*
$\hat{Q} = \{\hat{q}_1, \hat{q}_2, \dots\}$
Platform's trajectory depends on p, q, and $\hat{Q}$
$\hat{q}^\infty$
q
$\hat{q}'^{,\infty}$
$\hat{q}^0$
$\hat{Q}' = \{\hat{q}'_1, \hat{q}'_2, \dots\}$
$\hat{Q} = \{\hat{q}_1, \hat{q}_2, \dots\}$

Figure 2: Convergence of platform beliefs about the user as $t \to \infty$. (Left) Suppose the user adopts strategy q, and the platform begins with an initial belief μ_0. For illustrative purposes, we visualize the platform's initial belief using the corresponding estimate $\hat{q}^0$, and we use the orange polygon to represent ConvexHull$(\widehat{Q})$. As the platform collects data, its estimate evolves, eventually converging. We characterize the beliefs to which the platform converges using the *globally stable set*, as defined in Definition 4.2. In this figure, we visualize the stable set $\widehat{Q}_\infty \subset \widehat{Q}$ as a singleton set $\widehat{Q}_\infty = \{\hat{q}^\infty\}$, meaning that the platform's long-run belief will be the point-mass belief $\mu_\infty = \delta_{\hat{q}^\infty}$. (Right) As formalized in Definition 4.2, the belief to which the platform converges depends on the platform's strategy $(p, \widehat{Q})$ and the user's strategy q. We illustrate this dependence by visualizing how changing the platform's hypothesis class (from $\widehat{Q}$ to $\widehat{Q}'$) affects the platform's limiting belief (from $\delta_{\hat{q}^\infty}$ to $\delta_{\hat{q}',\infty}$).

4.2 Strategic user

In contrast to a naive user, a *strategic* user maximizes their *long-term expected payoff*. A strategic user does this by, first, considering how the platform's belief μ_t evolves as $t \to \infty$ if the user adopts some strategy q. The user uses this understanding to predict its payoff under all possible strategies q as $t \to \infty$, then selects a strategy that achieves the highest long-term payoff.

We formalize this idea in two stages. First, to characterizes the platform's long-term belief as a function of the user strategy q, we define the notion of a globally stable set.

Definition 4.2 (Globally stable set). *A set $\widehat{Q}_\infty \subset \widehat{Q}$ is $(q, p, \widehat{Q})$-globally stable under hypothesis class $\widehat{Q}$, algorithm p, and user strategy q if and only if, for any full-support initial belief μ_0,*

$$\mathbb{P}(\mu_t(\widehat{Q}_\infty) \to 1) = 1 \qquad \textit{as} \qquad t \to \infty,$$

where the probability above is taken with respect to the dynamics given in Section 3.

Intuitively, a set $\widehat{Q}_\infty \subset \widehat{Q}$ is *globally stable* if and only if it contains the support of the platform's limiting belief (i.e., of μ_t as $t \to \infty$) under strategies $(q, p, \widehat{Q})$. While the entire hypothesis class $\widehat{Q}$ is trivially a globally stable set, more fine-grained stable sets let us characterize how the platform behaves in the long run under a given $(q, p, \widehat{Q})$. In some cases, we will show the existence a globally stable *singleton* set, meaning that the platform's beliefs converge (almost surely) to a specific $\hat{q} \in \widehat{Q}$ that depends on $(q, p, \widehat{Q})$. Figure 2 illustrates this case.

Next, we define the platform and user's expected payoffs.

Definition 4.3 (Expected payoffs)**.** *Consider a distribution $r \in \Delta(\mathcal{Z})$ over propositions $\mathcal{Z}$ and a user strategy $q \in \mathcal{Q}$. Then, the platform's and user's expected payoffs under (r, q) are*

$$\overline{V}(r,q) := \mathbb{E}\left[V(Z,B)\right],$$
$$\overline{U}(r,q) := \mathbb{E}\left[U(Z,B) - \lambda \cdot d_{\mathcal{Q}}(q(\cdot|Z), q^{BR}(\cdot|Z))\right], \tag{2}$$

where the expectations are taken with respect to $Z \sim r$ and $B \sim q(\cdot|Z)$, $d_{\mathcal{Q}}$ is some distance metric over $\Delta(\mathcal{B})$, and $\lambda \geqslant 0$. The penalty term $\lambda \cdot d_{\mathcal{Q}}(q, q^{BR})$ in (2) captures the effort that the user expends to deviate from their naive (best-response) behavior.

Equipped with these definitions, we can now define the strategic user as a user who maximizes their expected payoff under the platform's worst-case, limiting behavior. Since different choices of $(q, p, \widehat{\mathcal{Q}})$ can induce different globally stable sets (Definition 4.2), we define the strategic user with respect to a *function* $S(q, p, \widehat{\mathcal{Q}})$ that maps $(q, p, \widehat{\mathcal{Q}})$ to a corresponding globally stable set.

Definition 4.4 (Strategic user)**.** *Let $S(q, p, \widehat{\mathcal{Q}})$ be a function that maps a user strategy q and platform strategy $(p, \widehat{\mathcal{Q}})$ to a $(q, p, \widehat{\mathcal{Q}})$-globally stable set $\widehat{\mathcal{Q}}_\infty$, as defined in Definition 4.2. Then, we define the S-strategic user as a user who adopts the strategy $q^\star_S(p, \widehat{\mathcal{Q}})$, where*

$$q^\star_S(p, \widehat{\mathcal{Q}}) \in \arg\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(S(q,p,\widehat{\mathcal{Q}}))} \overline{U}(p(\cdot;\mu), q). \tag{3}$$

To tease apart Definition 4.4, consider each component of (3). First, recall from Definition 4.2 that a set of user models $\widehat{\mathcal{Q}}$ is *globally stable* if it contains all the user models to which the platform assigns positive probability as $t \to \infty$. That is, if the user and platform adopt strategies $(q, p, \widehat{\mathcal{Q}})$ and S is as defined in Definition 4.4, the platform's beliefs as $t \to \infty$ are contained in $\Delta(S(q, p, \widehat{\mathcal{Q}}))$.

Second, note that $\overline{U}(p(\cdot;\mu), \cdot)$ is the user's expected payoff if the platform uses algorithm p to generate propositions under belief μ. Putting these two observations together,

$$\min_{\mu \in \Delta(S(q,p,\widehat{\mathcal{Q}}))} \overline{U}(p(\cdot;\mu), q)$$

is the S-strategic user's envisioned worst-case expected payoff as $t \to \infty$. The user then chooses the strategy $q \in \mathcal{Q}$ that maximizes this worst-case, limiting payoff. Therefore, a S-strategic user maximizes their worst-case limiting payoff under the chosen strategies $(q, p, \widehat{\mathcal{Q}})$ and mapping S.

Importantly, a strategic user pays attention to the platform's strategy $(p, \widehat{\mathcal{Q}})$ whereas a naive user's q^{BR} is the same regardless of the platform's chosen strategy. We illustrate the differences between naive and strategic users in Figure 3.

Remarks on globally stable sets. A few remarks are in order. First, there is no unique globally stable set, in general. As previously mentioned, for instance, $\widehat{\mathcal{Q}}$ is always a trivially globally stable set because $\mu_t(\widehat{\mathcal{Q}}) = 1$ for all t. Even so, we will show in Section 6 that there is a principled way to obtain rather fine-grained globally stable sets.

Second, when the globally stable set contains a single user model (i.e., $S^\infty(q, p, \widehat{\mathcal{Q}}) = \{\hat{q}^\infty\}$), we are guaranteed that the platform belief converges to a the point mass belief $\mu_\infty = \delta_{\hat{q}^\infty}$, and thus that the platform generates propositions from $p(\cdot;\mu_\infty)$. However, as we discuss later on, we cannot always guarantee that the platform converges to a single unique belief μ_∞. In these cases, a globally stable set characterizes the set of *possible* limiting beliefs.

Finally, even though strategization is defined with respect to a function S, when S is clear from context we omit it and say "strategic user" instead of "S-strategic user." Similarly, we omit S or $\widehat{\mathcal{Q}}$ from our notation for the strategic user's strategy $q^\star_S(p, \widehat{\mathcal{Q}})$ (see (3)) when clear from context.

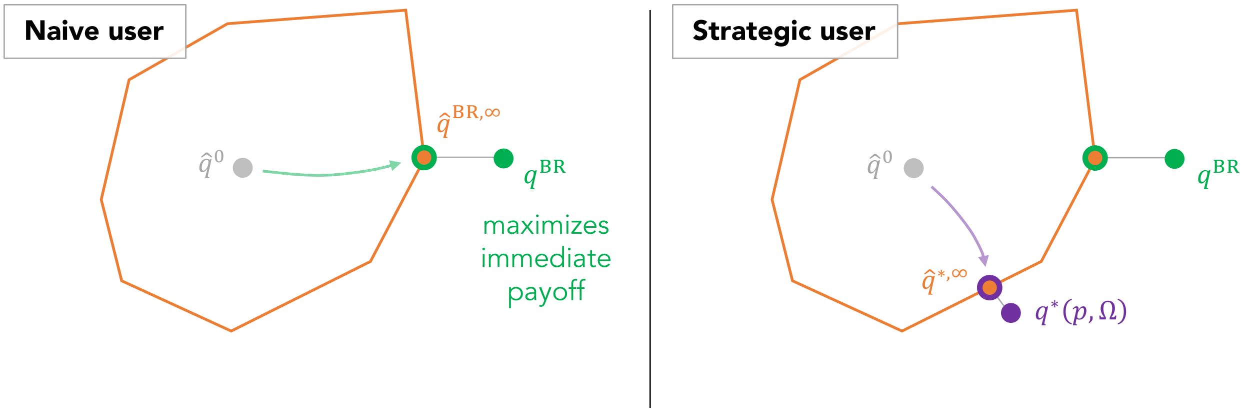


Figure 3: Illustration of a naive user (Section 4.1) on the left and a strategic user (Section 4.2) on the right. (Left) The (convex hull of the) platform's hypothesis class $\widehat{\mathcal{Q}}$ is given by the orange polygon. The naive user's strategy q^{BR} is given by the solid green dot. As in Figure 2, the platform's estimate of q^{BR} evolves as $t \to \infty$; we visualize the limiting estimate as $\hat{q}_i = \hat{q}^{\mathrm{BR},\infty}$. (Right) The strategic user considers their payoff under the platform's limiting estimate, i.e., $\overline{U}(p^{\delta_{\hat{q}_i}}, q^{\mathrm{BR}})$ and finds that they can instead adopt the strategy $q^*(p, \widehat{\mathcal{Q}})$ that leads the platform to a belief (and in turn, a proposition distribution) that is more favorable for the user.

5 Stylized Example

What are the implications of strategization? Is it good or bad for platforms? In this section, we consider a styled setting that allows us to answer these questions, and study how user strategization impacts a data-driven platform. Our goal is to illustrate and motivate our main results, which we establish in their full generality in Section 6.

At a high level, we will consider a simple recommender system that partitions its users into "types" (e.g., comedy lovers and horror lovers) and recommends content based on these types. The platform's payoff will be high when the user engages with the recommendations, and the user's payoff will depend on both their engagement and their personal taste. Within this setting, we demonstrate that:

1. The user is incentivized to be strategic—i.e., there is a user strategy that guarantees higher payoff than the naive (best-response) strategy q^{BR} defined in Definition 4.1. (Proposition 5.2)

2. Whenever the recommendation strategy $(p, \widehat{\mathcal{Q}})$ is fixed, the platform's payoff is *never lower* when the user behaves strategically than when the user is naive. (Proposition 5.3)

3. At the same time, user strategization can indirectly hurt the platform. Specifically, suppose that the platform wishes to update its recommendation algorithm from p to to a counterfactual algorithm p_{CF}. We show that the data that a platform obtains under p cannot be used to reliably make inferences under p_{CF}. (Proposition 5.4)

4. In the same vein, user strategization makes it harder to predict the effect of design choices. Specifically, we show that, when the user is strategic, expanding the hypothesis class $\widehat{\mathcal{Q}}$ can unexpectedly hurt the platform's payoff. That is, if the platform uses a richer class of user models to estimate the user's preferences, the platform's payoff can actually *decrease* when the user is strategic. (Proposition 5.5)

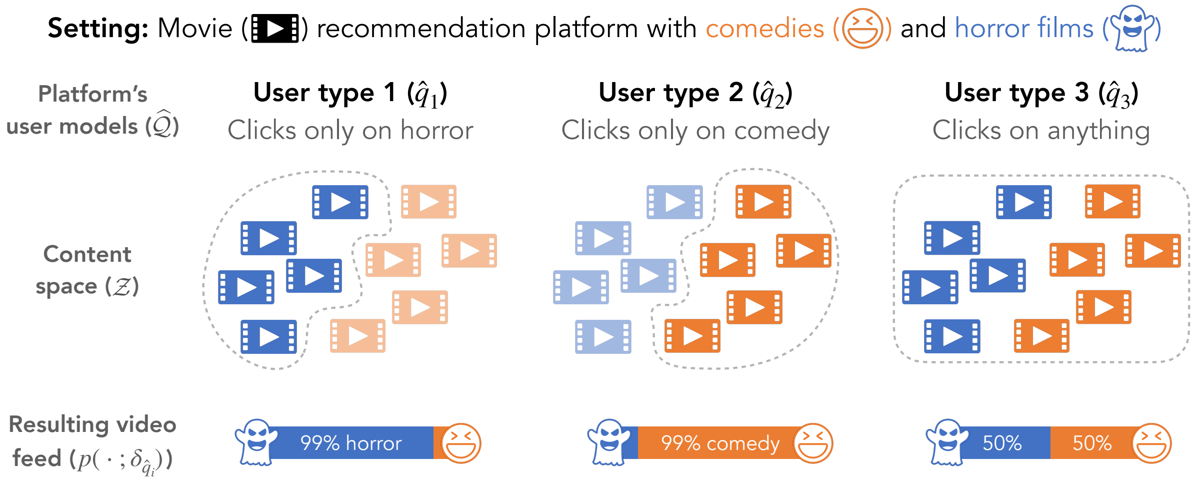

Setting: Movie (▶) recommendation platform with comedies (😆) and horror films (👻)
Platform's user models ($\widehat{\mathcal{Q}}$)
User type 1 ($\hat{q}_1$)
Clicks only on horror
User type 2 ($\hat{q}_2$)
Clicks only on comedy
User type 3 ($\hat{q}_3$)
Clicks on anything
Content space ($\mathcal{Z}$)
Resulting video feed ($p(\cdot\,; \delta_{\hat{q}_i})$)
99% horror
99% comedy
50%
50%

Figure 4: The recommender system that we consider in our stylized example. The platform's hypothesis class consists of three user models. Under one model, the user watches exclusively horror movies; under another, exclusively comedies; and under the last model, the user is equally interested in comedy and horror. The platform represents the user as a convex combination of these models, which will dictate the recommendations that the platform gives.

5.1 A simple recommender system

Consider a platform that recommends from a finite set of items $\mathcal{Z}$ and allows the user to click or ignore the recommended item (i.e., let $\mathcal{B} = \{0, 1\}$). Let the platform's and user's payoffs be

$$V(Z, B) = B, \qquad U(Z, B) = B \cdot a(Z), \tag{4}$$

for all $Z \in \mathcal{Z}$ and $B \in \mathcal{B}$, where $a : \mathcal{Z} \to \{-1, 1\}$ is a fixed function that encodes the user's affinity for item Z. Intuitively, the platform gains utility when the user clicks on any recommendation, and the user gets (possibly negative) utility $a(Z)$ from clicking on an item Z, and 0 from not clicking.

User types and the platform's hypothesis class. Suppose that there are two disjoint types of content on the platform, $\mathcal{Z}_1$ and $\mathcal{Z}_2$ (e.g., horror movies and comedies) with $\mathcal{Z}_2 = \mathcal{Z} \setminus \mathcal{Z}_1$. Let there be three types of users: those who prefer $\mathcal{Z}_1$ (i.e., $a(Z) \leqslant \mathbf{1}\{Z \in \mathcal{Z}_1\}$), those who prefer $\mathcal{Z}_2$ (i.e., $a(Z) \leqslant \mathbf{1}\{Z \in \mathcal{Z}_2\}$), and those who fall into neither of the two former categories. Note that $a(Z) \leqslant \mathbf{1}\{Z \in \mathcal{Z}_1\}$ means that the user definitely does not enjoy anything outside of $\mathcal{Z}_1$ and sometimes enjoys content in $\mathcal{Z}_1$.

The platform is aware of the three user types, but it does not know the correct partitioning $(\mathcal{Z}_1, \mathcal{Z}_2)$ and instead believes that the two kinds of content are $(\mathcal{Z}_A, \mathcal{Z}_B)$. This setting is common and can be generalized to capture instances in which the platform does not account for all possible users (e.g., there are "minority" users that do not follow mainstream trends).

The platform thus estimates user behavior using the hypothesis class $\widehat{\mathcal{Q}} = \{\hat{q}_1, \hat{q}_2, \hat{q}_3\}$, where

$$\hat{q}_i(B = 1 | Z) = \begin{cases} (1-\gamma)\mathbf{1}\{Z \in \mathcal{Z}_A\} & \text{if } i = 1, \\ (1-\gamma)\mathbf{1}\{Z \in \mathcal{Z}_B\} & \text{if } i = 2, \\ (1-\gamma)\mathbf{1}\{Z \in \mathcal{Z}\} & \text{if } i = 3 \end{cases} \qquad \forall\, Z \in \mathcal{Z}, \tag{5}$$

and $\gamma > 0$ is a constant that captures the fact that users will not click on *all* items of a given type.

Intuitively, the platform believes there are three possible users: users tend to like either only content in $\mathcal{Z}_A$, or only content in $\mathcal{Z}_B$, or all content. Formally, this means that under the user model $\hat{q}_1$ the user only clicks on items in $\mathcal{Z}_A$ and does so with probability $1 - \gamma$. Under $\hat{q}_2$, the user behaves analogously toward $\mathcal{Z}_B$. Under $\hat{q}_3$, the user clicks on any item with probability $1 - \gamma$.

Recommendation algorithm. Finally, suppose that the platform uses a simple algorithm p that with small probability ε (which we specify later) recommends a random item $Z \in \mathcal{Z}$, and otherwise recommends Z_i with probability proportional to its likelihood of inducing a "click" from the user (under the platform's current belief about the user). That is, for all $Z \in \mathcal{Z}$,

$$p(Z;\mu) := \underbrace{\varepsilon \cdot \frac{1}{|\mathcal{Z}|}}_{\text{Uniform w.p. } \varepsilon} + \underbrace{(1-\varepsilon) \cdot \frac{\sum_{\hat{q}_i \in \widehat{\mathcal{Q}}} \mu(\hat{q}_i) \cdot \hat{q}_i(B=1|Z)}{\sum_{Z \in \mathcal{Z}} \sum_{\hat{q}_i \in \widehat{\mathcal{Q}}} \mu(\hat{q}_i) \cdot \hat{q}_i(B=1|Z)}}_{\text{Proportional to click probability } \hat{q}(B=1|Z) \text{ w.p. } 1-\varepsilon}, \tag{6}$$

where recall from Section 3 that μ is the platform's belief, and so $\mu(\hat{q})$ is the probability that the platform assigns to the user's strategy being $\hat{q}$. We visualize this setup in Fig. 4.

5.2 User behavior

Now, consider a user of the first type, i.e., a user for which $a(Z) \leqslant \mathbf{1}\{Z \in \mathcal{Z}_1\}$. Define the set $\mathcal{Z}^+ = \{Z : a(Z) = 1\} \subset \mathcal{Z}_1$ as the items that the user enjoys. Recall that a user presented with a recommendation Z_t responds with behavior B_t sampled from their behavior strategy $q(\cdot|Z_t)$, where this behavior strategy depends on whether the user is *naive* or *strategic*, as follows.

Naive users. In this setting, a naive user will click on items for which they have positive affinity (i.e., for which $a(Z) = 1$), and ignore items for which they have negative affinity, i.e.,

$$q^{\mathrm{BR}}(B=1|Z) = \mathbf{1}\{a(Z) \geqslant 0\} \;\forall\, Z \in \mathcal{Z}. \tag{7}$$

Strategic users. A strategic user, on the other hand, chooses a strategy that elicits the highest long-term payoffs. For example, they might not click on an item Z that they like (i.e., for which $a(Z) = 1$), in order to influence the distribution of recommended items from the platform. To characterize the strategic user, we first need to understand the limiting beliefs of the platform.

Proposition 5.1. *Let $\widehat{\mathcal{Q}}$ be the hypothesis class defined by (5). Consider a user strategy q and a platform algorithm p such that $p(\cdot;\mu)$ has full support for all $\mu \in \Delta(\widehat{\mathcal{Q}})$. Let $supp(q) = \{Z \in \mathcal{Z} : q(B=1|Z) > 0\}$. If $|supp(q)| > 0$, then the following function S maps $(q, p, \widehat{\mathcal{Q}})$ to a globally stable set:*

$$S(q,p,\widehat{\mathcal{Q}}) := \{\hat{q}_{i^\star}\}, \qquad \textit{where} \qquad i^\star = \begin{cases} 1 & \textit{if } |supp(q) \cap \mathcal{Z}_B| = 0, \\ 2 & \textit{if } |supp(q) \cap \mathcal{Z}_A| = 0, \\ 3 & \textit{otherwise.} \end{cases}$$

In other words, the platform's limiting belief is $\mu_\infty = \delta_{\hat{q}_{i^\star}}$.

Proof. See Appendix A.1. □

Now, with this established, we show that unless the content that the user likes belongs entirely to $\mathcal{Z}_A$ or to $\mathcal{Z}_B$ (i.e., $\mathcal{Z}^+ \subset \mathcal{Z}_A$ or $\mathcal{Z}^+ \subset \mathcal{Z}_B$), then strategization *strictly* improves the user's utility.

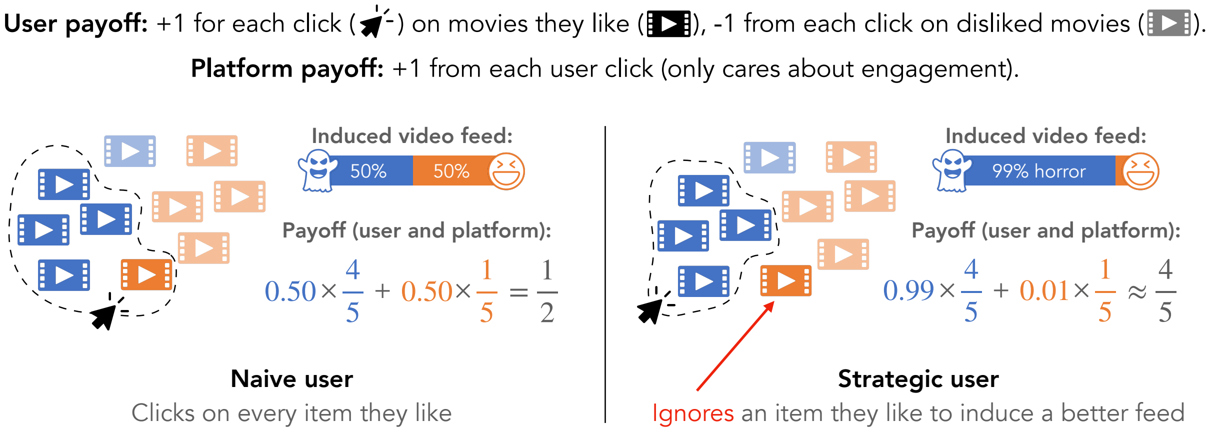

User payoff: +1 for each click ( ) on movies they like ( ), -1 from each click on disliked movies ( ).
Platform payoff: +1 from each user click (only cares about engagement).
Induced video feed:
50%
50%
Payoff (user and platform):
0.50 × 4/5 + 0.50 × 1/5 = 1/2
Naive user
Clicks on every item they like
Induced video feed:
99% horror
Payoff (user and platform):
0.99 × 4/5 + 0.01 × 1/5 ≈ 4/5
Strategic user
Ignores an item they like to induce a better feed

Figure 5: Naive and strategic user strategies in our stylized example (Section 5). We consider a user whose affinity $a(Z)$ is encoded by items' opacity in the Figure above. A naive strategy for this user (left) would click on item Z if and only if $a(Z) = 1$. This strategy would result in the platform modeling the user as the "clicks on anything" user $\hat{q}_3$ (see Fig. 4), and thus serve a feed that is 50% comedy and 50% horror. If the user is strategic (right), they recognize that the naive strategy is suboptimal, and they avoid clicking on "outlier" comedy videos that they enjoy. The platform thus estimates the user as a "clicks only on horror" user $\hat{q}_1$, and serves a feed that better suits the user. Notably, *both user and platform payoffs are higher* when the user is strategic.

Proposition 5.2. *Consider the setting described in Section 5.1, and suppose that the platform's partition $(\mathcal{Z}_A, \mathcal{Z}_B)$ is not "orthogonal" to the user's preferences, i.e.,*

$$\frac{|\mathcal{Z}^+ \cap \mathcal{Z}_A|}{|\mathcal{Z}_A|} \neq \frac{|\mathcal{Z}^+ \cap \mathcal{Z}_B|}{|\mathcal{Z}_B|}.$$

Then, for sufficiently small ε in (6), a strategic user's $q^\star = q^{BR}$ if and only if $\mathcal{Z}^+ \subset \mathcal{Z}_A$ or $\mathcal{Z}^+ \subset \mathcal{Z}_B$.

Proof. See Appendix A.2. Intuitively, if the user is naive and the platform is misspecified, the platform will learn $\mu = \delta_{\hat{q}_3}$ by Proposition 5.1, and will thus recommend a uniform distribution over all items. If one of $\mathcal{Z}_A$ or $\mathcal{Z}_B$ are closer to $\mathcal{Z}_1$, the strategic user can improve their utility by restricting their clicks to only the set $\mathcal{Z}^+ \cap \mathcal{Z}_A$ (which will lead to the platform recommending from $\mathcal{Z}_A$) or $\mathcal{Z}^+ \cap \mathcal{Z}_B$ (which will lead to the platform recommending from $\mathcal{Z}_B$). □

5.3 User strategization improves platform payoffs

We now study the impact of strategic behavior on the platform. We begin by showing that user strategization actually improves (or at least does not hurt) the platform's payoffs.

Proposition 5.3. *Consider the setting described in Section 5.1. The platform's payoff is as least as high when the user is strategic as when the user is naive.*

Proof. See Appendix A.3. Intuitively, the user only strategizes in order to get "better" content in the long term, and the platform benefits from this behavior because it receives a positive payoff every time the user clicks. □

Note that Proposition 5.3 holds regardless of the platform's choice of strategy $(p, \widehat{\mathcal{Q}})$, and depends only on the structure of the payoff functions U and V.

5.4 User strategization results in unexpected behavior

Now, suppose that, having deployed algorithm p and converged to the limiting belief μ_∞ given by Proposition 5.1, the platform considers changing its algorithm to downweight toxic content. That is, the platform considers replacing p with a counterfactual algorithm p_{CF}:

$$p_{\mathrm{CF}}(Z;\mu) \propto \textsc{toxicity}(Z) \cdot p(Z;\mu), \tag{8}$$

where for some $\alpha \in (0,1)$, the function $\textsc{toxicity}(Z) \in \{\alpha, 1\}$ discounts content Z that is toxic. We show that strategic behavior (on the part of the user) can *corrupt the platform's data* so that data gathered under p cannot be used to make reliable inferences under p_{CF}.

Unreliable counterfactual inferences. A natural way for the platform to gauge whether using the algorithm p_{CF} is a good idea is to *predict* its (counterfactual) payoff under $(p_{\mathrm{CF}}, \widehat{\mathcal{Q}})$. Specifically, the platform tries to estimate $\overline{V}^\star(p_{\mathrm{CF}}, \widehat{\mathcal{Q}})$, where for any p we define $\overline{V}^\star(p, \widehat{\mathcal{Q}})$ as

$$\overline{V}^\star(p, \widehat{\mathcal{Q}}) := \max_{\mu \in \Delta(S(q^\star, p, \widehat{\mathcal{Q}}))} \overline{V}(p(\cdot;\mu), q^\star(p, \widehat{\mathcal{Q}})), \tag{9}$$

and we recall that $q^\star(p, \widehat{\mathcal{Q}})$ is the strategic user's strategy in response to the platform strategy $(p, \widehat{\mathcal{Q}})$. Of course, the platform does not have access to $\overline{V}^\star(p_{\mathrm{CF}}, \widehat{\mathcal{Q}})$, as it does not know what the user's strategy *would be* in response to $(p_{\mathrm{CF}}, \widehat{\mathcal{Q}})$, and so it must instead predict its payoff using its current user model $\hat{q}_{i^\star}$ (collected under p). In other words, it computes

$$\widehat{V}(p_{\mathrm{CF}}, \widehat{\mathcal{Q}}) := \overline{V}(p_{\mathrm{CF}}(\cdot;\delta_{\hat{q}_{i^\star}}), \hat{q}_{i^\star}), \tag{10}$$

where we recall that $\hat{q}_{i^\star}$ is the platform's limiting belief from playing the algorithm p.

We show below that for some choices of the toxicity function, the above approach (i.e., (10)) will produce a drastically wrong estimate of the platform's long-run utility under p_{CF}. In particular, this mis-estimation might cause the platform to think that switching to p_{CF} will hurt utility when it will actually greatly help utility.

Proposition 5.4. *Consider the setting described in Section 5.1, and the counterfactual algorithm p_{CF} given by (8). For any content partitioning $(\mathcal{Z}_A, \mathcal{Z}_B)$ where $|\mathcal{Z}_A|, |\mathcal{Z}_B| \geqslant 4$, there exists an affinity function $a(Z)$ (see (4)), constants $\gamma > 0$ (see (5)) and $\varepsilon > 0$ (see (6)) and a function* $\textsc{toxicity} : \mathcal{Z} \to \{\alpha, 1\}$ *such that, by applying the strategy above:*

- *(a) the platform correctly predicts its own utility under p, i.e., $\widehat{V}(p, \widehat{\mathcal{Q}}) = \overline{V}^*(p, \widehat{\mathcal{Q}})$;*
- *(b) the platform thinks its payoff will decrease if it switches to algorithm p_{CF}, i.e., $\widehat{V}(p_{\mathrm{CF}}, \widehat{\mathcal{Q}}) < \overline{V}^*(p, \widehat{\mathcal{Q}})$;*
- *(c) in reality, the platform's payoff will increase if it switches to p_{CF}, i.e., $\overline{V}^\star(p_{\mathrm{CF}}, \widehat{\mathcal{Q}}) > \overline{V}^\star(p, \widehat{\mathcal{Q}})$.*

Proof. See Appendix A.4. The intuition here is that when the user strategizes and only engages with items from $\mathcal{Z}^+ \cap \mathcal{Z}_A$, the platform has no information about how much the user likes items from $\mathcal{Z}_B$. If, then, under p_{CF} the user switches to engaging only with items from $\mathcal{Z}^+ \cap \mathcal{Z}_B$ (say, due to elements in $\mathcal{Z}^+ \cap \mathcal{Z}_A$ being marked as toxic), the platform will not be able to predict its payoff accurately. □

We generalize (and in fact, strengthen) this result in Section 6 (Proposition 6.13).

Expanding $\widehat{\mathcal{Q}}$ can hurt the platform when users are strategic. To further illustrate the counter-intuitive phenomena that can occur when users are strategic, we also demonstrate that expanding the hypothesis class of user models can actually *hurt* the platform. (This finding is unexpected, as considering a richer class of models typically does not hurt a learning algorithm.)

Proposition 5.5. *Consider the setting described in Section 5.1. For any partitioning $(\mathcal{Z}_A, \mathcal{Z}_B)$ of $\mathcal{Z}$ there exists an affinity function $a(Z)$ (see (4)), constants $\gamma > 0$ (see (5)) and $\varepsilon > 0$ (see (6)), and a user model $\hat{q}_4$ such that when $\hat{q}_4$ is added to hypothesis class $\widehat{\mathcal{Q}}$, the platform's payoff under strategization decreases.*

Proof. See Appendix A.5. Intuitively, the platform can inadvertently eliminate the user's *means of strategization*, by adding a candidate user model $\hat{q}_4$ that is more similar to $q^\star$ than $\hat{q}_{i^\star}$ is to $q^\star$, but whose corresponding distribution over content $p(\cdot; \delta_{\hat{q}_4})$ is unfavorable for the user. This will incentivize the user switch to a strategy that lowers the platform's payoff. □

We prove a general version of this result in Section 6 (Proposition 6.14).

6 User strategization and its discontents

In this section, we present our main results. These results generalize and strengthen our findings from Section 5 and can be summarized as follows.

1. In Section 6.2, we characterize the platform's limiting belief about the user. In particular, we show that the platform's belief converges to the set of user models that are closest to the user's chosen behavior strategy q in an information-theoretic sense. This results allows us to characterize the platform's beliefs as $t \to \infty$ and, as a result, how the strategic user behaves.
2. We then show that user strategization can *help* the platform. Specifically, when the user and platform payoffs are aligned, then strategization improves both user and platform outcomes.
3. We also find that, although user strategization can improve outcomes under the platform's *current* strategy, user strategization can interfere with the platform's ability to anticipate how changes to their strategy $(p, \widehat{\mathcal{Q}})$ affect their payoff. In particular, when a user is strategic, (1) the data that a platform obtains under their current algorithm p cannot reliably predict the platform's payoff under a different algorithm p_{CF}; and (2) counter to what the platform might expect, expanding the hypothesis class $\widehat{\mathcal{Q}}$ can hurt its payoff.
4. We show that, on the other hand, if the user is naive, it is straightforward to anticipate how changing (p, Ω) affects the platform's payoff. This finding begs the question: When are users incentivized to play their best-response behaviors?

6.1 Preliminaries

We now introduce definitions and assumptions that we use in the remainder of the section.

Definitions. For any two probability measures Π_1 and Π_2 defined on a measurable space $(\Omega, \mathcal{F})$, with probability mass (or density) functions π_1 and π_2, the Kullback-Leibler divergence and total variation distance are given by

$$\mathrm{KL}(\Pi_1, \Pi_2) := \mathbb{E}_{x \sim \pi_1}\left[\log\left(\frac{\pi_1(x)}{\pi_2(x)}\right)\right] \quad \text{(defined when } \Pi_1 \text{ is absolutely continuous w.r.t. } \Pi_2\text{)},$$

$$\mathrm{TV}(\Pi_1, \Pi_2) := \sup_{A \in \mathcal{F}} |\Pi_1(A) - \Pi_2(A)|$$

Recall that $p : \Delta(\widehat{\mathcal{Q}}) \to \Delta(\mathcal{Z})$ denotes the platform's *algorithm*. We quantify the distance between two algorithms using their maximum total variation distance; that is, we let

$$d_{\mathcal{P}}(p_1, p_2) := \sup_{\mu \in \Delta(\widehat{\mathcal{Q}})} \mathrm{TV}(p_1(\cdot;\mu), p_2(\cdot;\mu)).$$

Recall that q denotes the user's strategy. Let $p(\cdot;\mu) \times q$ denote the joint distribution of (Z, B) when the platform's propositions are drawn according to $Z \sim p(\cdot;\mu)$ and the user's behavior is given by $B \sim q(\cdot|Z)$. Similarly, for any user model $\hat{q}_i \in \widehat{\mathcal{Q}}$, let $p(\cdot;\mu) \times \hat{q}_i$ denote the joint distribution of (Z, B) when $Z \sim p(\cdot;\mu)$ and the user's behavior is given by $B \sim \hat{q}_i(\cdot|Z)$.

For a fixed proposition distribution $p(\cdot;\mu)$ and a fixed user strategy q, we say that the user model $\hat{q}_i$ *strictly dominates* the user model $\hat{q}_j$ when $p(\cdot;\mu) \times \hat{q}_i$ better explains $p(\cdot;\mu) \times q$ than $p(\cdot;\mu) \times \hat{q}_j$ does (in the information-theoretic sense described below).

Definition 6.1 (Strict KL dominance)**.** *A user model $\hat{q}_i$ strictly KL-dominates $\hat{q}_j$ at $(p(\cdot;\mu), q)$, as denoted by $\hat{q}_i \succ^{q}_{p(\cdot;\mu)} \hat{q}_j$ if and only if $KL(p(\cdot;\mu) \times q, p(\cdot;\mu) \times \hat{q}_i) < KL(p(\cdot;\mu) \times q, p(\cdot;\mu) \times \hat{q}_j)$.*

Pervasive assumptions. We now lay out the assumptions that we will use in the remainder of this work. Throughout the entire section (even when not explicitly mentioned), we will make the following two assumptions adapted from Frick et al. [FII20].

Assumption 6.2 (Frick et al. [FII20])**.** The distributions $p(\cdot;\mu) \times q$ and $p(\cdot;\mu) \times \hat{q}_i$ are continuous Radon-Nikodym derivatives with respect to some σ-finite measure ν on $\mathcal{Z} \times \mathcal{B}$. When $\mathcal{Z}$ is discrete (resp., continuous), ν is a product of the counting (resp., Lebesgue) measure on $\mathcal{Z}$ and the counting measure on $\mathcal{B}$. In particular, $p(\cdot;\mu)$, $q(\cdot|Z)$, and $\hat{q}_i(\cdot|Z)$ are all well-defined probability densities.

Assumption 6.3 (Frick et al. [FII20])**.** The platform's recommendation algorithm p and hypothesis class $\widehat{\mathcal{Q}}$ satisfy the following three conditions:

1. (Support). For any user strategy q, user model $\hat{q}$, and $Z \in \mathcal{Z}$, supp $q(\cdot|Z) \subset$ supp $\hat{q}(\cdot|Z)$.
2. (Bounded likelihood ratios). There exists a ν-integrable function $h(Z, B)$ such that

$$\sup_{\mu \in \Delta(\widehat{\mathcal{Q}})} \max_{\hat{q}_1, \hat{q}_2 \in \widehat{\mathcal{Q}}} \frac{\hat{q}_1(B|Z)}{\hat{q}_2(B|Z)} \cdot p(Z;\mu) \cdot q(B|Z) \leqslant h(B, Z) \text{ for all } B, Z \in \mathcal{B} \times \mathcal{Z}.$$

3. (Belief continuity). For each user model $\hat{q} \in \widehat{\mathcal{Q}}$, there exists a neighborhood $\mathcal{N} \ni \delta_{\hat{q}}$ such that for all $Z \in \mathcal{Z}$ and $\mu \in \mathcal{N}$, the function $p(Z;\mu)$ is continuous in μ.

Assumption 6.2 simply ensures that the probability distributions we are dealing with are well-defined, while Assumption 6.3 establishes (mild) conditions under which we can characterize the platform's limiting belief (which the rest of our analysis relies on).

Regularity assumptions. In the coming sections, we will also make certain assumptions about the platform payoff, algorithm, and hypothesis class being well-behaved. Rather than being necessary for our negative results in the coming sections, these assumptions (Assumptions 6.4 to 6.6) actually strengthen our results. In particular, we will show (in Section 6.4) that *in spite of* these regularity conditions, platform payoffs are still poorly behaved and unpredictable when users are strategic. Unlike Assumptions 6.2 and 6.3, we explicitly reference the following assumptions when they are in place.

Assumption 6.4 (Payoff landscape)**.** For any distribution $r \in \Delta(\mathcal{Z})$ and user behavior q, and for any $\varepsilon > 0$, there exists $r' \in \Delta(\mathcal{Z})$ such that $\mathrm{TV}(r, r') < \varepsilon$ and $\overline{V}(r, q) \neq \overline{V}(r', q)$.

Assumption 6.5 (Well-behaved algorithm)**.** We call an algorithm $p : \Delta(\widehat{\mathcal{Q}}) \to \Delta(\mathcal{Z})$ well-behaved if it maps similar beliefs μ (in terms of the corresponding user models) to similar recommendation distributions $p(\cdot; \mu)$, i.e., if

$$d_{\mathcal{P}}(p(\cdot; \mu_1), p(\cdot; \mu_2)) \leqslant L_{\mathcal{P}} \cdot \mathbb{E}_{\hat{q}_1 \sim \mu_1, \hat{q}_2 \sim \mu_2}\left[\max_{Z \in \mathcal{Z}} \mathrm{TV}(\hat{q}_1(\cdot|Z), \hat{q}_2(\cdot|Z))\right] \qquad \forall\, \mu_1, \mu_2 \in \Delta(\widehat{\mathcal{Q}}).$$

Assumption 6.6 (Hypothesis class expansion)**.** A setup $(\mathcal{Z}, \mathcal{B}, V, p, \widehat{\mathcal{Q}})$ satisfies the *hypothesis class expansion* assumption if for any fixed user strategy q any two hypothesis classes $\widehat{\mathcal{Q}}_1, \widehat{\mathcal{Q}}_2 \subset \widehat{\mathcal{Q}}$ such that $\widehat{\mathcal{Q}}_1 \subset \widehat{\mathcal{Q}}_2$, and a globally stable set function S (as defined in Definition 4.2),

$$\min_{\mu \in S(q, p, \widehat{\mathcal{Q}}_1)} \overline{V}(p(\cdot; \mu), q) \leqslant \min_{\mu \in S(q, p, \widehat{\mathcal{Q}}_2)} \overline{V}(p(\cdot; \mu), q).$$

In other words, for a fixed user strategy q, the platform cannot decrease its payoff by expanding its hypothesis class.

6.2 Platforms converge to beliefs that best approximate user in KL-sense

We can now begin characterizing user strategization. Recall that strategic users plan ahead—they are aware that their behavior q affects the platform's future propositions, so the strategic user chooses q in order to obtain high *long-term* payoffs. How do users anticipate what will happen in the long-term? In this section, we show that platforms converge to beliefs that best approximate the user's behavior q in an information-theoretic sense, which allows users to reason about their long-term payoffs.

Drawing from a long line of work [EP16; Boh16; FLS21b; FII20], we use a concept introduced by Frick et al. [FII20] known as the *iterated elimination of dominated states*, wherein, given a subset $\widehat{\mathcal{Q}}' \subset \widehat{\mathcal{Q}}$, one repeatedly eliminates user models $\hat{q}_j$ that are strictly KL-dominated by another user model $\hat{q}_i$ in the set (according to Definition 6.1).

Definition 6.7 (Iterated elimination of dominated parameters)**.** *Consider a platform strategy* $(p, \widehat{\mathcal{Q}})$ *and user strategy* q*. Define the elimination operator* $R : 2^{\widehat{\mathcal{Q}}} \to 2^{\widehat{\mathcal{Q}}}$ *as*

$$R(\widehat{\mathcal{Q}}') = \{\hat{q}_j \in \widehat{\mathcal{Q}}' : \nexists\, \hat{q}_i \in \widehat{\mathcal{Q}}' \text{ such that } \hat{q}_i \succ^{q}_{p(\cdot;\mu)} \hat{q}_j \;\forall\, \mu \in \Delta(\widehat{\mathcal{Q}}')\}.$$

Let $R^0(\widehat{\mathcal{Q}}) = \widehat{\mathcal{Q}}$ *and recursively define* $R^n(\widehat{\mathcal{Q}}) = R(R^{n-1}(\widehat{\mathcal{Q}}))$*. Finally, define* $S^\infty(q, p, \widehat{\mathcal{Q}}) = \cap_{n=1}^{\infty} R^n(\widehat{\mathcal{Q}})$*.*

Frick et al. [FII20] show that the fixed point of the elimination operator $S^\infty(q, p, \widehat{\mathcal{Q}})$ is a globally stable set, i.e., the platform converges to the parameters that best approximate q in a KL-sense.

Theorem 6.8 (Theorem [FII20])**.** *For algorithm p and user strategy q, $S^{\infty}_{p,q}(\Omega)$ is (p,q)-globally stable.*

Theorem 6.8 allows us to characterize strategic users. Suppose, for example, that for a given user strategy q, $S^{\infty}(q, p, \widehat{\mathcal{Q}})$ contains a single element $\hat{q}^*(q)$. In this case, we can be certain that the platform's belief converges to $\delta_{\hat{q}^*(q)}$ as $t \to \infty$. A S^{∞}-strategic user would then choose the strategy q that maximizes their limiting payoff $\overline{U}(p(\cdot; \delta_{\hat{q}^*(q)}), q)$.

There are many cases in which $S^{\infty}(q, p, \widehat{\mathcal{Q}})$ reduces to a single element, known as the *Berk-Nash equilibrium* [EP16]. In general, however, $S^{\infty}(q, p, \widehat{\mathcal{Q}})$ can contain more than one element. In these cases, Theorem 6.8 tells us that the platform's proposition distribution as $t \to \infty$ is contained within the set $\{p(\cdot; \mu) : \mu \in \Delta(S^{\infty}(q, p, \widehat{\mathcal{Q}}))\}$.

6.3 User strategization can help the platform

Our first main result shows that user strategization can *improve* the platform's payoff $\overline{V}$. This occurs when the payoffs of the user and platform are sufficiently aligned.

To see why this might be the case, consider YouTube. YouTube would like to engage and retain users (by serving users good recommendations), while also ensuring the profitability of their platform. Although users on Youtube may not care about the platform's profitability, they do want good recommendations. In this way, there is some alignment between the user and platform payoff. Users on YouTube often have an idea of how the YouTube algorithm works and, in response, adapt to the algorithm in order to elicit better recommendations. This strategization can ultimately help the platform by improving their recommendations and therefore increasing user engagement.

The following result (whose interpretation we discuss below) corroborates this relationship, showing that user strategization can increase the platform's payoffs.

Proposition 6.9. *Consider a platform strategy $(p, \widehat{\mathcal{Q}})$ and suppose that $U(B, Z)$ has a unique maximizer in $\mathcal{B}$ for all Z. For a user strategy q, let $\widetilde{V}(q)$ be the platform's worst-case limiting payoff,*

$$\widetilde{V}(q) := \min_{\mu \in \Delta(S^{\infty}(q,p,\widehat{\mathcal{Q}}))} \overline{V}(p(\cdot; \mu), q), \qquad \forall q \in \mathcal{Q}, \tag{11}$$

where S^{∞} is defined in Definition 6.7. Then, user strategization strictly improves the platform's worst-case limiting payoff if

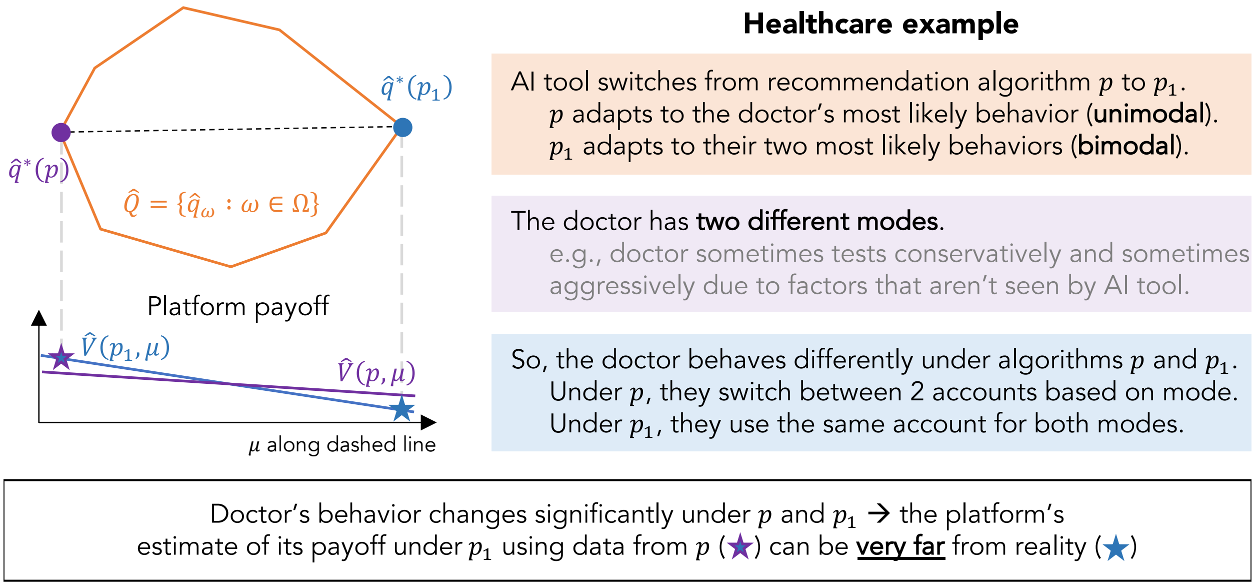
$$\widetilde{V}\left(\arg\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(S^{\infty}(q,p,\widehat{\mathcal{Q}}))} \overline{U}(p(\cdot; \mu), q)\right) > \widetilde{V}\left(\arg\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(\widehat{\mathcal{Q}})} \overline{U}(p(\cdot; \mu), q)\right). \tag{12}$$

The same is true if the min *in* (11) *is swapped out for a* max*.*

Proposition 6.9 follows almost by definition (see Appendix B.1), and gives a simple condition under which the platform's payoff is *strictly higher* when a user strategizes than when a user is naive. In particular, there are two takeaways from Proposition 6.9:

1. *Strategization helps platforms when U and V are sufficiently aligned.* Observe that the right-hand side of (11) is identical to the analogous expression on the left-hand side of (12), except that $\overline{V}$ is swapped out for $\overline{U}$ in (12). Therefore, when U and V are sufficiently aligned, finding a q that maximizes the worst-case user payoff (i.e., $\min_{\mu \in \Delta(S^{\infty}(q,p,\widehat{\mathcal{Q}}))} \overline{U}(p(\cdot; \mu), q)$) will also yield a high platform payoff.

   The following corollary confirms this intuition, showing that user strategization can never lower the platform's payoff when U and V are perfectly aligned.



Healthcare example

AI tool switches from recommendation algorithm p to p_1.
p adapts to the doctor's most likely behavior (**unimodal**).
p_1 adapts to their two most likely behaviors (**bimodal**).

The doctor has **two different modes**.
e.g., doctor sometimes tests conservatively and sometimes aggressively due to factors that aren't seen by AI tool.

So, the doctor behaves differently under algorithms p and p_1.
Under p, they switch between 2 accounts based on mode.
Under p_1, they use the same account for both modes.

Doctor's behavior changes significantly under p and p_1 → the platform's estimate of its payoff under p_1 using data from p (★) can be **very far** from reality (★)

Figure 6: A platform cannot estimate its counterfactual payoff well when a user is strategic. That is, the platform's estimated payoff under an algorithm p_{CF} if the data used to estimate its payoff is collected under p can be arbitrarily far from the true value (Proposition 6.13). On the left, we visualize how, even when p and p_{CF} are close and V is smooth in its first argument, the platform can misestimate its payoff under p_{CF}. In the proof of Proposition 6.13, we provide intuition for settings in which $\hat{q}^*(p)$ and $\hat{q}^*(p_{\text{CF}})$ can be far apart even when p and p_{CF} are close.

Corollary 6.10. *The platform's payoff under a strategic user is at least as high as its payoff under a naive user when $U = V$.*

2. *Strategization can be viewed as a form of coordination.* This follows from the fact that the left- and right-hand sides of (12) look similar, with the only difference being the set of beliefs over which μ is minimized. The set on the left-hand side is constrained because the user anticipates how the platform will behave. When the user and platform have aligned incentives, and the user has a good idea of how the platform will behave, the resulting "coordination" helps both the user and the platform.

   The following corollary solidifies this intuition, showing that even when user and platform have only partially aligned payoffs, they can coordinate (via strategization) to find a region of proposition space where their payoffs are aligned.

Corollary 6.11. *The platform's payoff under a strategic user is at least as high as its payoff under a naive user when there exist functions $g : \mathcal{B} \to \mathbb{R}$ and $f : \mathcal{Z} \to \{-1, 1\}$ such that user and platform payoffs decompose as $U(B, Z) = g(B) \cdot f(Z)$ and $V(B, Z) = g(B)$ respectively.*

6.4 User strategization can cause unexpected behavior

Although user strategization can help the platform under its current strategy $(p, \widehat{\mathcal{Q}})$, we now show that strategization interferes with learning by "corrupting" the data that the platform collects. We further show strategization can cause other unexpected behavior—specifically, expanding the hypothesis class $\widehat{\mathcal{Q}}$ that the platform uses can unexpectedly *hurt* the platform under strategization.

6.4.1 Changing the algorithm p results in unpredictable payoff

In this section, consider a fixed hypothesis class $\widehat{\mathcal{Q}}$. Suppose that, after deploying algorithm p, the platform wants to change its algorithm to a counterfactual algorithm p_{CF}. In these cases, the platform may wish to *estimate* their expected payoff if they were to switch to p_{CF}. Formally, for any belief $\mu \in \Delta(\widehat{\mathcal{Q}})$ and algorithm p', we define the platform's *predicted payoff* as

$$\widehat{V}(p', \mu) := \mathbb{E}_{\hat{q}\sim\mu}\left[\overline{V}(p'(\cdot;\mu), \hat{q})\right]. \tag{13}$$

Now, consider a user who is S^{∞}-strategic. Recall from Section 4.2 that $q^{\star}(p_{\mathrm{CF}}, \widehat{\mathcal{Q}})$ denotes the strategy that the strategic user adopts if the platform employs strategy $(p_{\mathrm{CF}}, \widehat{\mathcal{Q}})$. We can consequently write the platform's worst-case payoff under p_{CF} and user strategization as

$$\overline{V}^{\star}(p_{\mathrm{CF}}) := \min_{\mu\in\Delta(S^{\infty}(q,p_{\mathrm{CF}}))} \overline{V}(p_{\mathrm{CF}}, q^{\star}(p_{\mathrm{CF}})). \tag{14}$$

The following result shows that if a platform gathers data under p but is interested in estimating its payoff under an alternative algorithm p_{CF}, its estimate may be arbitrarily bad when the user is strategic. This result holds *even when the platform is (nearly) correctly specified*, i.e., even for the highly expressive hypothesis class $\widehat{\mathcal{Q}}$ defined below:

Definition 6.12 (ε-net hypothesis class)**.** *For a finite proposition space $\mathcal{Z}$ and some $\varepsilon > 0$, let $\Delta_{\varepsilon}(\mathcal{B})$ be an ε-net of $\Delta(\mathcal{B})$ with respect to the ℓ_{∞} metric. An ε-net hypothesis class is the set of all possible mappings from propositions in $\mathcal{Z}$ to behavior distributions in $\Delta_{\varepsilon}(\mathcal{B})$, i.e., $\widehat{\mathcal{Q}}_{\varepsilon} := \Delta_{\varepsilon}(\mathcal{B})^{\mathcal{Z}}$ (and thus $|\widehat{\mathcal{Q}}_{\varepsilon}| \approx (\frac{1}{\varepsilon})^{|B|\cdot|Z|}$).*

By proving the result below under Definition 6.12, we rule out the case where the platform is unable to predict counterfactual payoffs simply because it cannot accurately model the user.

Proposition 6.13. *Consider a given platform strategy $(p, \widehat{\mathcal{Q}})$ and platform payoff function V. Suppose that $\widehat{\mathcal{Q}}$ is an ε-net (Definition 6.12) for some sufficiently small ε, that $p(\cdot;\mu)$ has full support for all μ, and that Assumption 6.4 holds. Define ζ as the maximum gap in predicted platform payoff, i.e.,*

$$\zeta(p') = \max_{\hat{q}_1,\hat{q}_2\in\mathcal{Q}} \overline{V}(p'(\cdot;\delta_{\hat{q}_1}), \hat{q}_1) - \overline{V}(p'(\cdot;\delta_{\hat{q}_2}), \hat{q}_2).$$

Further assume that there exists $\beta \geqslant \alpha \geqslant 0$ such that $\alpha \leqslant Var_{p(\cdot;\mu)\times q}[V(B,Z)] \leqslant \beta$ for any belief μ and user strategy q. Then, for any $\varepsilon_0, \varepsilon_1 > 0$, there exists a p_{CF} and U such that $d_{\mathcal{P}}(p, p_{\mathrm{CF}}) \leqslant \varepsilon_0$, and and

$$\min_{\mu\in S^{\infty}(q,p)} \left|\widehat{V}(p_{\mathrm{CF}}, \mu) - \overline{V}^{\star}(p_{\mathrm{CF}})\right| \geqslant \sqrt{\zeta(p_{\mathrm{CF}})^2 - 4(\beta-\alpha)} - \varepsilon_1. \tag{15}$$

Proof. See Appendix B.2. The intuition behind the proof is given in Figure 6. □

If the variance of the platform's payoff does not change much across beliefs (i.e., if $\alpha \approx \beta$), the right side of (15) is effectively the largest gap that can exist between an estimated payoff and the true payoff. As such, Proposition 6.13 states that data collected under p can be unhelpful for making predictions about a counterfactual algorithm (see Fig. 6 for an example). The intuition behind this result is that, when users strategize, the platform limiting payoff is *nonsmooth* in p, i.e., there exists a p_{CF} that is ε-close to p for which the payoff under strategization is far from that under p.

6.4.2 Expanding $\widehat{\mathcal{Q}}$ can unexpectedly hurt the platform

Next, we show that when a user is strategic, expanding $\widehat{\mathcal{Q}}$ can hurt the platform's payoff. This is somewhat counterintuitive, as $\widehat{\mathcal{Q}}$ is the hypothesis class that the platform users to infer the user's behavior q, and expanding one's model family typically does not hurt estimation.

To make this statement formal, suppose that the user is S^∞-strategic and that the algorithm p is fixed. (Since p is held fixed in this section, we suppress its notation below.) Let $q^\star(\widehat{\mathcal{Q}})$ denote the strategy that the strategic user adopts if the platform uses hypothesis class $\widehat{\mathcal{Q}}$. That is, let

$$q^\star(\widehat{\mathcal{Q}}) := \arg\max_{q \in \mathcal{Q}} \min_{\mu \in S^\infty(q, \widehat{\mathcal{Q}})} \overline{U}(p(\cdot;\mu), q).$$

Proposition 6.14. *Consider a platform strategy $(p, \widehat{\mathcal{Q}})$ and a platform payoff function V bounded in $[0,1]$ (without loss of generality). Suppose $(\mathcal{Z}, \mathcal{B}, V, p, \widehat{\mathcal{Q}})$ satisfy the expansion assumption (Assumption 6.6), and that $V(Z,B)$ has a unique maximizer with respect to B for each $Z \in \mathcal{Z}$. Then, there exist a user payoff function U and sets $\widehat{\mathcal{Q}}_1, \widehat{\mathcal{Q}}_2 \subset \widehat{\mathcal{Q}}$ such that $\widehat{\mathcal{Q}}_1 \subseteq \widehat{\mathcal{Q}}_2$, but*

$$\min_{\mu \in S^\infty(q^\star(\widehat{\mathcal{Q}}_1), \widehat{\mathcal{Q}}_1)} \overline{V}(p(\cdot;\mu), q^\star(\widehat{\mathcal{Q}}_1)) > \max_{\mu \in S^\infty(q^\star(\widehat{\mathcal{Q}}_2), \widehat{\mathcal{Q}}_2)} \overline{V}(p(\cdot;\mu), q^\star(\widehat{\mathcal{Q}}_2)).$$

Proof. See Appendix B.3. The intuition is that the platform can unintentionally remove the user's *means of strategization*. That is, a user may want to induce a specific belief from the platform without straying too far from their best-response strategy. Thus, they purposefully pick a strategy q_1^* that the platform misinterprets in a desirable way. When the platform gets "better" at capturing their behavior by adding a user model that is close to q_1^*, the user is forced to move even further away from their best-response behavior, making things worse for the platform. □

Like Proposition 6.13, Proposition 6.14 shows that user strategization can cause unexpected behavior. Both imply that strategization makes it difficult to use its data under one strategy $(p, \widehat{\mathcal{Q}})$ make inferences about a different strategy.

6.5 System behavior is more predictable for best-response users

The previous section establishes that strategization can make it difficult for the platform to predict how the system behaves if either p or $\widehat{\mathcal{Q}}$ are changed. This is consequential because the platform may wish to use the data that they have collected under $(p, \widehat{\mathcal{Q}})$ to estimate some quantity under a counterfactual $(p', \widehat{\mathcal{Q}}')$. In this section, we show that *these problems do not arise when users are naive*. In other words, when the user plays according to their best-response at each timestep, the platform benefits because it is easier for them to make inferences under changes to $(p, \widehat{\mathcal{Q}})$.

Formally, suppose we have a best-response user, and fix a platform hypothesis class $\widehat{\mathcal{Q}}$. Recall the platform's payoff estimator (13) (restated below), and let $\overline{V}_{\mathrm{BR}}(p_{\mathrm{CF}})$ denote the true payoff under $(p_{\mathrm{CF}}, \widehat{\mathcal{Q}})$ (i.e., analagous to (14)) when the user is naive:

$$\begin{aligned}
\widehat{V}(p_{\mathrm{CF}}, \mu) &= \mathbb{E}_{\hat{q} \sim \mu}\left[\overline{V}(p_{\mathrm{CF}}(\cdot;\mu), \hat{q})\right]. \\
\overline{V}_{\mathrm{BR}}(p_{\mathrm{CF}}) &= \min_{\mu \in \Delta(S^\infty(q^{\mathrm{BR}}, p_{\mathrm{CF}}))} \overline{V}\left(p_{\mathrm{CF}}(\cdot;\mu), q^{\mathrm{BR}}\right).
\end{aligned}$$

Our first result shows that, in contrast to Proposition 6.13, using data gathered under p to estimate the platform's payoff under p_{CF} is a good idea for best-response users.

Proposition 6.15. *Suppose that the platform's strategy $(p, \widehat{\mathcal{Q}})$ is such that the hypothesis class $\widehat{\mathcal{Q}}$ is an ε-net (Definition 6.12). Let p_{CF} be a counterfactual algorithm that is well-behaved (Assumption 6.5); then,*

$$\max_{\mu \in \Delta(S^{\infty}(q^{BR}, p))} \left| \widehat{V}(p_{CF}, \mu) - \overline{V}_{BR}(p_{CF}) \right| \leqslant \sqrt{\varepsilon} \left((2L_{\mathcal{P}} + 1) \sqrt{|\mathcal{B}|} \right). \tag{16}$$

As a result, by using a sufficiently fine ε-net hypothesis class $\widehat{\mathcal{Q}}$, the platform can estimate its payoff under counterfactual algorithms up to arbitrary precision.

Proof. See Appendix B.4. □

Next, suppose that the hypothesis class $\widehat{Q}$ can change but the algorithm p is fixed. In addition, the next result shows that, in direct contrast to Proposition 6.14, payoffs cannot decrease when the hypothesis class $\widehat{\mathcal{Q}}$ is expanded and the user plays naively.

Proposition 6.16. *Consider a given $(p, \widehat{\mathcal{Q}})$ and V. Under Assumption 6.6, for any $\widehat{\mathcal{Q}}' \subset \widehat{\mathcal{Q}}$, Then*

$$V(\widehat{\mathcal{Q}}, q^{BR}) \geqslant V(\widehat{\mathcal{Q}}', q^{BR}).$$

Proof. This follows directly from Assumption 6.6. □

Propositions 6.15 and 6.16 show that incentivizing the user to play their best-response behavior can lead to more reliable data and payoffs for the platform, begging the question: When are users incentivized to play naively?

7 Trustworthy algorithm design

In the previous sections, we found that strategic behavior can hurt the platform. Specifically, we showed that user strategization violates a key assumption of most data-driven algorithms that user behavior is exogeneous, which compromises the platform's ability to re-use the data that it collects, e.g., to train future algorithms. In contrast, we found that the platform's data *looks* exogenous if the user behaves naively. That is, a platform can recover the exogeneity assumption if it encourages users to behave naively.

In this section, we argue that this analysis suggest that *trustworthy design* can help platforms. We discuss how trustworthiness produces two beneficial outcomes: users are not incentivized to strategize—which allows platforms to recover the exogeneity assumption—and users are incentivized to engage with the platform over alternatives. We begin in Section 7.1 with a formal definition of trustworthiness. Importantly, we draw a distinction between trustworthiness and strategy-proofness, connecting our definition to the concept of individual rationality in mechanism design. We then discuss how this definition relates to existing notions of trust in Section 7.2. In Sections 7.3-7.4, we use this definition to unpack four reasons why users do not trust their platforms and propose two interventions for building user trust.

7.1 Trustworthiness and its effect on the platform

Building on our analysis in Section 6, we present a formal definition of trustworthiness.

Definition 7.1 (Trustworthy). *Let $q^{\star}$ denote the policy that the S^{∞}_{Ω}-strategic user adopts when the platform employs algorithm p. A platform's policy $(p, \widehat{\mathcal{Q}})$ is κ-trustworthy when the user is not incentivized to strategize and the naive user's limiting payoff under $(p, \widehat{\mathcal{Q}})$ is at least $\kappa \in \mathbb{R}$, i.e.,*

1. $\min_{\mu \in \Delta(S^{\infty}_{\Omega}(p,q^{\star}))} \overline{U}(p^{\mu}, q^{\star}) \leqslant \min_{\mu \in \Delta(S^{\infty}_{\Omega}(p,q^{BR}))} \overline{U}(p^{\mu}, q^{BR})$ *and*
2. $\min_{\mu \in \Delta(S^{\infty}_{\Omega}(p,q^{BR}))} \overline{U}(p^{\mu}, q^{BR}) \geqslant \kappa$*.*

First requirement of trustworthiness. One can think of Definition 7.1(1) as follows: a platform satisfies the first requirement of trustworthiness if the user does not have to strategize because the platform looks out for the user's interests so they do not have to do so themselves. When the platform meets this requirement, the user might as well play their best-response action.

As an example, consider YouTube. Suppose that a user likes to alternate between "junk" content and "healthy" content based on their mood [KMR22]. If they were "truthful," then they would always pick the video that fits their current mood.[3] Suppose that the platform adopts an algorithm that only accounts for one possible mood and can therefore only model the user as liking "junk" or "healthy" content, but not both. Then, the user is better off behaving strategically; for instance, only using YouTube for one—but not both—of their moods (or maintaining two YouTube accounts, as has been observed anecdotally). Intuitively, the user does not *trust* that the platform interprets their behavioral data correctly. That is, they do not trust that the platform will use their behavioral data to generate good content in the future. On the other hand, if YouTube is able to correctly parse the user's mood and recommend content to suit both moods, the user does not have to behave strategically. The user therefore *trusts* the platform to correctly interpret their actions. A platform may also be deemed trustworthy (or untrustworthy) for many other reasons, such as how they protect user privacy.

Second requirement of trustworthiness. Definition 7.1(2) states that trustworthiness is earned only if the user's expected limiting payoff is at least κ when the user plays their best-response action. That is, it is not enough that the platform is strategy-proof, as required by Definition 7.1(1). Trustworthiness additionally requires that behaving naively is sufficiently beneficial for the user. Consider, for instance, a platform that adopts a strategy $(p, \widehat{\mathcal{Q}})$ under which the user's limiting payoff is -1 no matter what strategy the user adopts. Then, $(p, \widehat{\mathcal{Q}})$ satisfies Definition 7.1(1) because the user is not incentivized to strategize since all strategies induce a payoff of -1, but the system is *not* κ-trustworthy under Definition 7.1(2) for any $\kappa > -1$.

This requirement echoes *individual rationality*, a concept in mechanism design under which agents continue engaging with the platform (which is known in mechanism design as "participating in the mechanism") if it is beneficial for them to do so. Using this interpretation, κ determines the expected limiting payoff at which the user does not trust the platform although they may be willing to continue engaging with the platform. In this way, κ captures the user's ability to tolerate the untrustworthy behavior. The higher κ is, the more trust the user places in the platform, and the more likely the user is to engage with the platform over alternatives.

Implication of trustworthiness on the platform. As shown in Sections 5-6, trust is central to the platform's ability to collect reliable data. Without trust, users are incentivized to strategize, which is particularly harmful for platforms because it means that the data that they use for multiple purposes—such as training future algorithms or predicting the performance of candidate algorithms—is unreliable. More broadly, earning user trust is often beneficial to platforms when they rely not only on the continued participation of their users, but also on the amount of user engagement. The more time users spend on the platform, the more data the platform collects. This

[3] Note that this setting can be modeled by allowing U to be stochastic, with a hidden variable that represents the user's underlying stochastic mood.

facet of trustworthiness is captured by the second requirement in Definition 7.1—if a platform is only κ-trustworthy, but another platform is κ'-trustworthy for $\kappa' > \kappa$, users may be compelled to spend more time on the other platform (or even switch platforms). There are, of course, occasions when a platform is not incentivized to build trust, which could occur when the platform benefits so greatly from strategization (enough to overshadow the potential harms of unreliable data) that platform does not rely greatly on data collection for prediction, or there is little risk of users leaving the platform.

7.2 Elements of trustworthiness

The goal of this section is to place our definition of trustworthiness within the broader discussion of trustworthiness. Below, we identify several key elements of trustworthiness that appear in the literature on trust [Kra99; Har06; Nis01], where we examine situations in which "I" trust "you."

1. Trustworthiness does not imply that you and I have perfectly aligned interests; only that your behavior takes some of my interests into account.

2. Distrust arises when I expect to incur losses from interacting with you (unless I behave strategically) and, conversely, trust arises when I expect to gain from interaction.

3. Trust is inherently relational in that trusting you to look after my interests may depend on who "you" and "I" are, so that "you" are not necessarily universally trustworthy.

4. Trust is generally meaningful only when there are repeated interactions. In particular, I only put trust in you if I must rely on you in the future. Moreover, I only trust you to take my interests into account if I believe you value a continued relationship with me.

5. Trust involves vulnerability—the possibility of harm—because to trust is to allow another to affect one's interests.

All of these elements are also present in our Definition 7.1. First, a user and platform need not have the exact same interests for the results in Section 6, which show that untrustworthiness can be harmful to the platform, to hold. Second, distrust in our work arises when a user is incentivized to strategize because the platform does not account for the user's interests, meaning that trust and distrust are linked to gains and losses from interactions. Third, our formulation of trust is relational; a user's best-response and strategic behavior are specific to them, meaning that a platform's trustworthiness may differ across users. Fourth, a continuing relationship is pivotal to our formulation of trust. In fact it matters in two ways: (i) users only strategize because they can anticipate future interactions, and (ii) platforms typically rely on the continued participation of users for success. Lastly, user payoffs, as modeled in Section 3.1, depend on both the user's action as well as the platform's. As such, the user's happiness is influenced by the platform. This point is salient because the risk of being harmed is why trust often matters in data-driven decision-making.

Remark 7.2. *Note that there are other factors that may influence trust, such as the credibility of the trusted party, their reputation, and even whether they are virtuous or reliable. These factors are based on the perceived qualities of the trusted party. Although perception has a significant impact on trust, we leave perception-based factors trust to future discussions, focusing instead on settings in which trust relies primarily on other factors. By putting perception aside, we align more closely with conceptualizations of trust such as Hardin's, who states that trust is determined by the interests of each party and the desire of one party to take the other party's interests into account in order to foster a continuing relationship [Har06].*

7.3 Reasons why a user may not trust their platform

Under Definition 7.1 of trustworthiness, there are three reasons why a user might not trust their platform's policy. We discuss these three reasons in this section, then use these insights in the following section to inform how platforms can build user trustworthiness.

Misspecification. We say that a platform is *misspecified* with respect to q^{BR} when $q^{\mathrm{BR}} \notin \widehat{\mathcal{Q}}$. Misspecification implies that, should a user choose to play their best-response actions, the platform would not be able to model their behavior perfectly. For example, misspecification occurs when a platform believes that there are a few canonical "types" of users, but the user of interest does not fit into any of these types. Misspecification often induces users to strategize because the platform's inability to model the user correctly can mean that a user gains more by strategizing (e.g., by pretending to be one of the canonical types of user).

Hidden, changing state. Suppose that a user has different "modes." For instance, a user on an online shopping platform sometimes needs clothes for everyday wear and sometimes needs clothes for special occasions. Alternatively, suppose that there is salient information that is only available to the user but not to the platform.

In such cases, there is a state that changes across time and is hidden from the platform. We do not explicitly model this setting in our setup in Section 3; however, one can encompass such cases by adding a Step 0 to the game in Section 3.1 during which a state x_t is drawn at the start of each time step, i.e., "nature" selects a state x_t. The user then chooses action $B_t \sim q(\cdot|Z_t, x_t)$ in Step 2, and the user receives a payoff $U(Z_t, B_t, x_t)$ in Step 3. Because the platform does not have access to salient information x_t, the user may find that their x_t-dependent behavior is misattributed to other factors. This misattribution can lead the platform to behave unexpectedly (to the user's detriment), therefore prompting the user to strategize.

Algorithm incompatible with user payoffs. Recall that p denotes the platform's algorithm and U denotes the user's payoff. A user may strategize if p is incompatible with U. Even if the platform is not misspecified with respect to the best-response strategy q^{BR} and the platform has access to all salient information (such that the two reasons for strategization given above are absent), a user may wish to strategize if the platform's method for generating propositions is detrimental to a user following q^{BR}. This would, for instance, be the reason why users often strategize when an algorithm is known to have feedback loops or why users strategize when their recommendation algorithm shows too much content of type X after a user clicks on a piece of content of type X.

7.4 Interventions for improving trustworthiness

Before we discuss interventions for increasing trustworthiness under Definition 7.1, we first describe and explain why naive interventions that do not build trustworthiness but attempt to overcome the challenges untrustworthiness often fall short. We then discuss how two interventions for improving trustworthiness can complement these efforts.

Naive interventions. Recall that the main challenge of untrustworthiness is that user strategization distorts the platform's data, which compromises its ability to estimate user behavior under counterfactual strategies $(p, \widehat{\mathcal{Q}})$. The platform might wish to overcome the challenges of untrustworthiness by: (i) designing a strategy-proof mechanism, (ii) modeling the user's payoff function

U in order to predict their behavior under any counterfactual $(p, \widehat{\mathcal{Q}})$, (iii) expanding the hypothesis class $\widehat{\mathcal{Q}}$, (iv) guessing the user's hidden state x_t, as described in Section 7.3, or (v) improving one's algorithm p. Below, we discuss how these interventions can fall short of methods aimed at directly boosting trustworthiness.

First, designing strategy-proof mechanisms does not necessarily elicit to higher payoffs for the user or platform. For instance, an algorithm p that recommends YouTube videos that all users universally dislike is strategy-proof. In this scenario, the user's best-response (i.e., naive) strategy is not to click on any recommended video. This strategy is also their highest-payoff strategy because clicking on any video incurs a negative payoff. Therefore, p is strategy-proof, but it does not lead to positive outcomes for the user or the platform.

The second option seeks to model each user's payoff function U so that it is possible to infer how the user would behave under alternative platform strategies $(p, \widehat{\mathcal{Q}})$. Developing such models for complex settings is extremely difficult. In particular, this approach becomes challenging when users are heterogeneous, there is unobserved confounding, and the space of possible platform strategies $(p, \widehat{\mathcal{Q}})$ is large. Imagine, for instance, predicting how an arbitrary social media user would behave under any possible feed. Such an estimation task is notoriously challenging. For one, both the user's and platforms action spaces are large (i.e., the user can interact with content in many possible ways, and there are many possible posts that the platform could recommend). For another, each social media user behaves differently (i.e., there is a high heterogeneity). In similarly complex (or high-risk) settings, developing reliable models is difficult.

Third, one might naturally think to expand $\widehat{\mathcal{Q}}$ in order to address issues that arise from misspecification. We show in Proposition 6.14, however, that expanding the hypothesis class $\widehat{\mathcal{Q}}$ can,in fact, lower the platform's payoff and does not necessarily remove misspecification unless q^{BR} is guaranteed to be in the new hypothesis class.

Finally, we argue in the remainder of this section that the fourth and fifth approaches described above can indeed improve outcomes for the user and platform, but are less straightforward than interventions that directly boost trustworthiness, as given next.

Recommendation #1: Offering multiple algorithms. Offering users multiple algorithms from which they can choose addresses several issues simultaneously. As an example, consider Twitter, which offers personalized, chronological, and trending feeds. Returning to the three reasons users strategize, as given in Section 7.3, allowing users to select between multiple algorithms at each time step t is a straightforward way of accommodating unobserved confounding, e.g., a hidden state x_t. While the platform could predict x_t, doing so is inevitably less reliable that giving users the ability to select an algorithm based on x_t themselves. Second, if one algorithm is incompatible with a user's payoffs, the user will simply ignore that algorithm. On the other hand, if the platform only offers one algorithm p, it inevitably alienates users whose payoffs are incompatible with p.

Offering multiple algorithms can therefore diminish two of the reasons that users strategize (as given in Section 7.3) and guarantee that a user's limiting payoff is at least as high as under a single algorithm. It therefore improves trustworthiness by Definition 7.1.

Recommendation #2: Providing feedback mechanisms. Another intervention that builds trust and therefore mitigates the risks of strategization is providing users opportunities to give meaningful feedback. Returning to the reasons for strategization in Section 7.3, feedback can serve as a simple and reliable indicator for misspecification. If a user consistently indicates that the platform is not behaving in their interest, then the platform not only learns that this user is misspecified under $\widehat{\mathcal{Q}}$, but also gains insight into how $\widehat{\mathcal{Q}}$ can be improved. One can also view feedback as a "high-friction" or "high-cost" signal. Typically, signals with high friction or cost are more informative than low-friction or low-cost ones.

There are many ways of eliciting feedback, and not all mechanisms are made equal. Platforms must ensure that they are comprehensive but not overwhelming, easily accessible but not so pervasive that users simply treat it as an annoyance. Determining the precise design of feedback mechanisms is out of the scope of this work and may be of interest in future work.

References

[Abe+19] Jacob D Abernethy, Rachel Cummings, Bhuvesh Kumar, Sam Taggart, and Jamie H Morgenstern. "Learning auctions with robust incentive guarantees". In: *Advances in neural information processing systems* 32 (2019).

[ARS13] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. "Learning prices for repeated auctions with strategic buyers". In: *Advances in neural information processing systems* 26 (2013).

[Ban+19] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. "Beyond accuracy: The role of mental models in human-AI team performance". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), pp. 2–11.

[Bao+21] Ying Bao, Xusen Cheng, Triparna De Vreede, and Gert-Jan De Vreede. "Investigating the relationship between AI and trust in human-AI collaboration". In: *Hawaii International Conference on System Sciences 2021 (HICSS-54)*. aisel.aisnet.org, 2021.

[BBD08] Lucian Busoniu, Robert Babuska, and Bart De Schutter. "A comprehensive survey of multiagent reinforcement learning". In: *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society* 38.2 (Mar. 2008), pp. 156–172.

[BH21] J Aislinn Bohren and Daniel N Hauser. "Learning with heterogeneous misspecified models: Characterization and robustness". en. In: *Econometrica: journal of the Econometric Society* 89.6 (2021), pp. 3025–3077.

[BHK22] Gavin Brown, Shlomi Hod, and Iden Kalemaj. "Performative Prediction in a Stateful World". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by Gustau Camps-Valls, Francisco J R Ruiz, and Isabel Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, 2022, pp. 6045–6061.

[BKS12] Michael Brückner, Christian Kanzow, and Tobias Scheffer. *Static prediction games for adversarial learning problems*. https://www.jmlr.org/papers/volume13/brueckner12a/brueckner12a.pdf. Accessed: 2023-10-31. 2012.

[BM93] James Bergin and W Bentley MacLeod. "Continuous Time Repeated Games". In: *International economic review* 34.1 (1993), pp. 21–37.

[Boh16] J Aislinn Bohren. "Informational herding with model misspecification". In: *Journal of Economic Theory* 163 (2016), pp. 222–247.

[CIM22] S H Cen, A Ilyas, and A Madry. "A Game-Theoretic Perspective on Trust in Recommendation". In: *NeurIPS Workshop on Responsible Decision Making in Dynamic Environments (RDD)*. 2022.

[Den+21] Yuan Deng, Sebastien Lahaie, Vahab Mirrokni, and Song Zuo. "Revenue-Incentive Tradeoffs in Dynamic Reserve Pricing". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2601–2610.

[Don+18] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. "Strategic Classification from Revealed Preferences". In: *Proceedings of the 2018 ACM Conference on Economics and Computation*. EC '18. Ithaca, NY, USA: Association for Computing Machinery, June 2018, pp. 55–70.

[EP16] Ignacio Esponda and Demian Pouzo. "Berk–Nash equilibrium: A framework for modeling agents with misspecified models". In: *Econometrica* 84.3 (2016), pp. 1093–1130.

[Eze+19] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J Hepenstal, Christopher A Miller, and Dylan D Schmorrow. "Trust Engineering for Human-AI Teams". In: *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting Human Factors and Ergonomics Society. Meeting* 63.1 (Nov. 2019), pp. 322–326.

[FII20] Mira Frick, Ryota Iijima, and Yuhta Ishii. "Stability and robustness in misspecified learning models". In: (2020).

[FL93] Drew Fudenberg and David K Levine. "Self-Confirming Equilibrium". In: *Econometrica: journal of the Econometric Society* 61.3 (1993), pp. 523–545.

[FLS21a] Drew Fudenberg, Giacomo Lanzani, and Philipp Strack. "Limit points of endogenous misspecified learning". en. In: *Econometrica: journal of the Econometric Society* 89.3 (2021), pp. 1065–1098.

[FLS21b] Drew Fudenberg, Giacomo Lanzani, and Philipp Strack. "Limit points of endogenous misspecified learning". In: *Econometrica* 89.3 (2021), pp. 1065–1098.

[FRS17] D Fudenberg, G Romanyuk, and P Strack. "Active learning with a misspecified prior". In: *Theoretical Economics* (2017).

[Gha+21] Ganesh Ghalme, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. "Strategic Classification in the Dark". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3672–3681.

[Had+16] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. "Cooperative inverse reinforcement learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., Dec. 2016, pp. 3916–3924.

[Hag+22] Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. "Learning in Stackelberg Games with Non-myopic Agents". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Boulder, CO, USA: Association for Computing Machinery, July 2022, pp. 917–918.

[Har+16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. "Strategic Classification". In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*. ITCS '16. Cambridge, Massachusetts, USA: Association for Computing Machinery, Jan. 2016, pp. 111–122.

[Har06] Russell Hardin. *Trust*. Vol. 10. Polity, 2006.

[HHC23] Keke Hou, Tingting Hou, and Lili Cai. "Exploring Trust in Human–AI Collaboration in the Context of Multiplayer Online Games". en. In: *Systems* 11.5 (Apr. 2023), p. 217.

[HHP23] Andreas Haupt, Dylan Hadfield-Menell, and Chara Podimata. "Recommending to Strategic Users". In: *Foundations of Responsible Computing*. 2023.

[KL23] Hankyung Kim and Youn-Kyung Lim. "Investigating How Users Design Everyday Intelligent Systems in Use". In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. DIS '23. Pittsburgh, PA, USA: Association for Computing Machinery, July 2023, pp. 702–711.

[KL93] Ehud Kalai and Ehud Lehrer. "Subjective Equilibrium in Repeated Games". In: *Econometrica: journal of the Econometric Society* 61.5 (1993), pp. 1231–1240.

[KL95] Ehud Kalai and Ehud Lehrer. "Subjective games and equilibria". In: *Games and economic behavior* 8.1 (Jan. 1995), pp. 123–163.

[KMR22] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. EC '22. Boulder, CO, USA: Association for Computing Machinery, July 2022, p. 29.

[KN19] Yash Kanoria and Hamid Nazerzadeh. "Incentive-Compatible Learning of Reserve Prices for Repeated Auctions". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco, USA: Association for Computing Machinery, May 2019, pp. 932–933.

[Kra99] Roderick M Kramer. "Trust and distrust in organizations: Emerging perspectives, enduring questions". In: *Annual review of psychology* 50.1 (1999), pp. 569–598.

[LR22] Sagi Levanon and Nir Rosenfeld. "Generalized Strategic Classification and the Case of Aligned Incentives". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 12593–12618.

[Mar20] Aarian Marshall. *Uber Changes Its Rules, and Drivers Adjust Their Strategies*. Ed. by www.wired.com. Feb. 2020. URL: https://www.wired.com/story/uber-changes-rules-drivers-adjust-strategies/.

[Moz+22] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. "Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming". In: (Oct. 2022). arXiv: 2210.14306 [cs.SE].

[MS06] George J Mailath and Larry Samuelson. *Repeated games and reputations: Long-Run relationships*. New York, NY: Oxford University Press, July 2006.

[MT07] Frank McSherry and Kunal Talwar. "Mechanism Design via Differential Privacy". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. ieeexplore.ieee.org, Oct. 2007, pp. 94–103.

[Ned+22] Thomas Nedelec, Clément Calauzènes, Noureddine El Karoui, and Vianney Perchet. "Learning in Repeated Auctions". In: *Foundations and Trends® in Machine Learning* 15.3 (2022), pp. 176–334.

[Nis01] Helen Nissenbaum. "Securing trust online: Wisdom or oxymoron". In: *BUL Rev.* 81 (2001), p. 635.

[NST12] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. "Approximately optimal mechanism design via differential privacy". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: Association for Computing Machinery, Jan. 2012, pp. 203–213.

[OY20] Kazuo Okamura and Seiji Yamada. "Adaptive trust calibration for human-AI collaboration". en. In: *PloS one* 15.2 (Feb. 2020), e0229132.

[Per+20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. "Performative Prediction". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé Iii and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 7599–7609.

[Rot+10] Aaron Roth, Maria Florina Balcan, Adam Kalai, and Yishay Mansour. "On the Equilibria of Alternating Move Games". In: *Proceedings of the 2010 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Proceedings. Society for Industrial and Applied Mathematics, Jan. 2010, pp. 805–816.

[SHS22] Ellen Simpson, Andrew Hamann, and Bryan Semaan. "How to Tame "Your" Algorithm: LGBTQ+ Users' Domestication of TikTok". In: *Proc. ACM Hum.-Comput. Interact.* 6.GROUP (Jan. 2022), pp. 1–27.

[SVM22] Ignacio Siles, Luciana Valerio-Alfaro, and Ariana Meléndez-Moran. "Learning to like TikTok . . . and not: Algorithm awareness as process". In: *New Media & Society* (Dec. 2022), p. 14614448221138973.

[Tan93] Ming Tan. "Multi-agent reinforcement learning: Independent vs. cooperative agents". In: *Proceedings of the tenth international conference on machine learning*. books.google.com, 1993, pp. 330–337.

[Wan+19] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. "Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI". In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019), pp. 1–24.

[Wan+20] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. "From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA '20. Honolulu, HI, USA: Association for Computing Machinery, Apr. 2020, pp. 1–6.

[WAO20] M Wu, S Amin, and A Ozdaglar. "Multi-agent Bayesian Learning with Adaptive Strategies: Convergence and Stability". In: *arXiv preprint arXiv:2010.09128* (2020).

[WAO21] M Wu, S Amin, and A Ozdaglar. "Multi-agent Bayesian Learning with Best Response Dynamics: Convergence and Stability". In: *arXiv preprint arXiv:2109.00719* (2021).

[YSE19] Lantao Yu, Jiaming Song, and Stefano Ermon. "Multi-Agent Adversarial Inverse Reinforcement Learning". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 7194–7201.

[Zha+21] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. ""An Ideal Human": Expectations of AI Teammates in Human-AI Teaming". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW3 (Jan. 2021), pp. 1–25.

[Zrn+21] Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. “Who Leads and Who Follows in Strategic Classification?” In: *Advances in neural information processing systems* 34 (2021), pp. 15257–15269.

[ZYB21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. “Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms”. In: *Handbook of Reinforcement Learning and Control*. Ed. by Kyriakos G Vamvoudakis, Yan Wan, Frank L Lewis, and Derya Cansever. Cham: Springer International Publishing, 2021, pp. 321–384.

Appendix

Table of Contents

A Omitted Proofs: Illustrative Example

In this section, we provide the proofs omitted from Section 5.

A.1 Proof of Proposition 5.1

Proposition 5.1. *Let $\widehat{\mathcal{Q}}$ be the hypothesis class defined by (5). Consider a user strategy q and a platform algorithm p such that $p(\cdot;\mu)$ has full support for all $\mu \in \Delta(\widehat{\mathcal{Q}})$. Let $supp(q) = \{Z \in \mathcal{Z} : q(B = 1|Z) > 0\}$. If $|supp(q)| > 0$, then the following function S maps $(q, p, \widehat{\mathcal{Q}})$ to a globally stable set:*

$$S(q, p, \widehat{\mathcal{Q}}) := \{\hat{q}_{i^\star}\}, \qquad \textit{where} \qquad i^\star = \begin{cases} 1 & \textit{if } |supp(q) \cap \mathcal{Z}_B| = 0, \\ 2 & \textit{if } |supp(q) \cap \mathcal{Z}_A| = 0, \\ 3 & \textit{otherwise.} \end{cases}$$

In other words, the platform's limiting belief is $\mu_\infty = \delta_{\hat{q}_{i^\star}}$.

Proof. In this proof, we will use the result of Frick et al. [FII20], restated as Theorem 6.8. Note that the theorem cannot be applied directly due to the "bounded likelihood ratios" regularity condition being violated—still, since $p(\cdot;\mu)$ has full support, we can always wait until the violating user models are eliminated by Bayes' rule (which happens almost surely), and then apply the results from [FII20] thereafter.

For example, note that $\hat{q}_3(B|Z) > 0$ for all $B \in \mathcal{B}$ and $Z \in \mathcal{Z}$, so if $\mathrm{supp}(q) \cap \mathcal{Z}_A \neq \emptyset$, then with probability one we will eventually see a (B, Z) such that $B = 1$ and $Z \in \mathcal{Z}_A$, at which point

$$\mu(\hat{q}_2) = \frac{\mu_{t-1}(\hat{q}_2) \cdot \hat{q}_2(B_t|Z_t)}{\sum_{\hat{q}_i \in \widehat{\mathcal{Q}}} \mu_{t-1}(\hat{q}_i) \cdot \hat{q}_i(B_t|Z_t)} = 0,$$

and thus we can apply the results from [FII20] to the set $\{\hat{q}_1, \hat{q}_3\}$.

note that for *any* belief μ and any user model $\hat{q}$,

$$\begin{aligned} KL(p^\mu \times q, p^\mu \times \hat{q}) &= \mathbb{E}_{(B,Z)\sim p^\mu \times q}\left[\log\left(\frac{p^\mu(Z) \cdot q(B|Z)}{p^\mu(Z) \cdot \hat{q}(B|Z)}\right)\right] \\ &= \mathbb{E}_{Z\sim p^\mu}\left[\mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{q(B|Z)}{\hat{q}(B|Z)}\right)\right]\right], \end{aligned}$$

and so for the user model $\hat{q}_3$, $\varepsilon \leqslant \hat{q}_3(B|Z) \leqslant 1 - \varepsilon$ for all $B \in \mathcal{B}, Z \in \mathcal{Z}$, and so the quantity above is guaranteed to be finite.

Now, since the algorithm p_0 recommends a random item Z with probability ε,

$$\begin{aligned} KL(p^\mu \times q, p^\mu \times \hat{q}) &\geqslant \varepsilon \cdot \frac{1}{|\mathcal{Z}|} \sum_{Z\in\mathcal{Z}} \mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{q(B|Z)}{\hat{q}(B|Z)}\right)\right] \\ &= \varepsilon \cdot \frac{1}{|\mathcal{Z}|}\left(\sum_{Z\in\mathcal{Z}_A} \mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{q(B|Z)}{\hat{q}(B|Z)}\right)\right] + \sum_{Z\in\mathcal{Z}_B} \mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{q(B|Z)}{\hat{q}(B|Z)}\right)\right]\right) \\ &\geqslant \varepsilon \cdot \frac{1}{|\mathcal{Z}|}\left(\max_{Z\in\mathcal{Z}_A} q(B=1|Z) \cdot \log\left(\frac{q(B=1|Z)}{\hat{q}(B=1|Z)}\right)\right. \\ &\qquad \left. + \max_{Z\in\mathcal{Z}_B} q(B=1|Z) \cdot \log\left(\frac{q(B|Z)}{\hat{q}(B|Z)}\right)\right). \end{aligned}$$

Now, if $|\text{supp}(q) \cap \mathcal{Z}_A| > 0$, then the first maximum is infinite for $\hat{q} = \hat{q}_2$, since there will be at least one element $Z \in \mathcal{Z}_A$ for which $q(B = 1|Z) > 0$, but $\hat{q}_2(B = 1|Z) = 0$ for $Z \in \mathcal{Z}_A$. Similarly, if $|\text{supp}(q) \cap \mathcal{Z}_B| > 0$, the second maximum will be infinity for $\hat{q}_1$. This observation suffices to show that if both $|\text{supp}(q) \cap \mathcal{Z}_B| > 0$ and $|\text{supp}(q) \cap \mathcal{Z}_B| > 0$, the platform will converge to $\hat{q}_3$.

To complete the proof, suppose without loss of generality that $|\text{supp}(q) \cap \mathcal{Z}_B| = 0$. We need to prove that in this case, the platform will converge to $\hat{q}_1$ (and in particular, not $\hat{q}_3$). We will show this by using KL-dominance—in particular, for any belief μ,

$$\begin{aligned} KL(p^\mu \times q, p^\mu \times \hat{q}_3) - KL(p^\mu \times q, p^\mu \times \hat{q}_1) &= \mathbb{E}_{(B,Z)\sim p^\mu \times q}\left[\log\left(\frac{q(B|Z)}{\hat{q}_3(B|Z)}\right) - \log\left(\frac{q(B|Z)}{\hat{q}_1(B|Z)}\right)\right] \\ &= \mathbb{E}_{(B,Z)\sim p^\mu \times q}\left[\log\left(\frac{\hat{q}_1(B|Z)}{\hat{q}_3(B|Z)}\right)\right] \\ &= \mathbb{E}_{(B,Z)\sim p^\mu \times q}\left[\log\left(\frac{\hat{q}_1(B|Z)}{\hat{q}_3(B|Z)}\right)\right]. \end{aligned}$$

Since $\hat{q}_3(\cdot|Z) = \hat{q}_1(\cdot|Z)$ for $Z \in \mathcal{Z}_A$, this simplifies to

$$= \mathbb{E}_{Z\sim p^\mu}\left[\mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{\hat{q}_1(B|Z)}{\hat{q}_3(B|Z)}\right)\right] \middle| Z \in \mathcal{Z}_B\right] \cdot \mathbb{P}\left(Z \in \mathcal{Z}_B\right).$$

By assumption (i.e., that $|\text{supp}(q) \cap \mathcal{Z}_B| = 0$), we know that $q(\cdot|Z) = \delta_{B=0}$, and so

$$\begin{aligned} &= \mathbb{E}_{Z\sim p^\mu}\left[\log\left(\frac{\hat{q}_1(B = 0|Z)}{\hat{q}_3(B = 0|Z)}\right) \middle| Z \in \mathcal{Z}_B\right] \cdot \mathbb{P}\left(Z \in \mathcal{Z}_B\right) \\ &= \log\left(\frac{1}{\gamma}\right) \cdot \mathbb{P}\left(Z \in \mathcal{Z}_B\right) \\ &> 0. \end{aligned}$$

Thus, $\hat{q}_1$ strictly dominates $\hat{q}_3$ at all beliefs μ, and so the platform will converge to $\hat{q}_1$. □

A.2 Proof of Proposition 5.2

Proposition 5.2. *Consider the setting described in Section 5.1, and suppose that the platform's partition $(\mathcal{Z}_A, \mathcal{Z}_B)$ is not "orthogonal" to the user's preferences, i.e.,*

$$\frac{|\mathcal{Z}^+ \cap \mathcal{Z}_A|}{|\mathcal{Z}_A|} \neq \frac{|\mathcal{Z}^+ \cap \mathcal{Z}_B|}{|\mathcal{Z}_B|}.$$

Then, for sufficiently small ε in (6), a strategic user's $q^\star = q^{BR}$ if and only if $\mathcal{Z}^+ \subset \mathcal{Z}_A$ or $\mathcal{Z}^+ \subset \mathcal{Z}_B$.

Proof. We begin with the reverse direction. Suppose $\mathcal{Z}^+$ is not fully contained in either $\mathcal{Z}_A$ or $\mathcal{Z}_B$. By Proposition 5.1. If the user is naive, the platform will converge to the belief $\mu(\hat{q}_3) = 1$, and so the limiting proposition distribution will be a uniform distribution over all items (since $\hat{q}_3(B = 1|Z) = 1 - \varepsilon$ for all Z). The expected user payoff will then be equal to

$$\overline{U}(p^{\delta_{\hat{q}_3}}, q^{\text{BR}}) = \frac{|\mathcal{Z}^+|}{|\mathcal{Z}|}.$$

Without loss of generality, suppose $\mathcal{Z}_A$ is a better approximation than $\mathcal{Z}_B$ to $\mathcal{Z}_1$, and in particular

$$\Delta := \frac{|\mathcal{Z}_A \cap \mathcal{Z}^+|}{|\mathcal{Z}_A|} - \frac{|\mathcal{Z}_B \cap \mathcal{Z}^+|}{|\mathcal{Z}_B|} > 0.$$

We define ε (the probability with which p_0 recommends items uniformly at random) as

$$\varepsilon < \frac{|\mathcal{Z}_A| \cdot |\mathcal{Z}_B| \cdot \Delta}{|\mathcal{Z}_A \cap \mathcal{Z}^+|}.$$

Now, rearranging this definition yields

$$\begin{aligned}
|\mathcal{Z}_A| \cdot |\mathcal{Z}_B| \cdot \left(\frac{|\mathcal{Z}_A \cap \mathcal{Z}^+|}{|\mathcal{Z}_A|} - \frac{|\mathcal{Z}_B \cap \mathcal{Z}^+|}{|\mathcal{Z}_B|} \right) &> |\mathcal{Z}_A \cap \mathcal{Z}^+| \cdot \varepsilon \\
|\mathcal{Z}_B| \cdot |\mathcal{Z}_A \cap \mathcal{Z}^+| - |\mathcal{Z}_A| \cdot |\mathcal{Z}_B \cap \mathcal{Z}^+| &> |\mathcal{Z}_A \cap \mathcal{Z}^+| \cdot \varepsilon \\
(1-\varepsilon) \cdot |\mathcal{Z}_B| \cdot |\mathcal{Z}_A \cap \mathcal{Z}^+| &> \varepsilon \cdot |\mathcal{Z}_A \cap \mathcal{Z}^+| \cdot |\mathcal{Z}_A| + |\mathcal{Z}_A| \cdot |\mathcal{Z}_B \cap \mathcal{Z}^+| \\
(1-\varepsilon) \cdot |\mathcal{Z}_A \cap \mathcal{Z}^+| \cdot |\mathcal{Z}_A| + (1-\varepsilon) \cdot |\mathcal{Z}_B| \cdot |\mathcal{Z}_A \cap \mathcal{Z}^+| &> |\mathcal{Z}_A \cap \mathcal{Z}^+| \cdot |\mathcal{Z}_A| + |\mathcal{Z}_A| \cdot |\mathcal{Z}_B \cap \mathcal{Z}^+| \\
(1-\varepsilon) \cdot |\mathcal{Z}_A \cap \mathcal{Z}^+| \cdot |\mathcal{Z}| &> |\mathcal{Z}^+| \cdot |\mathcal{Z}_A| \\
\frac{|\mathcal{Z}_A \cap \mathcal{Z}^+|}{|\mathcal{Z}_A|} &> \frac{|\mathcal{Z}^+|}{|\mathcal{Z}|} \cdot \frac{1}{1-\varepsilon}.
\end{aligned}$$

Now, consider an alternative strategy q' where the user clicks only on items in $\mathcal{Z}_+ \cap \mathcal{Z}_A$. By Proposition 5.1, this would lead the platform to converge to $\hat{q}_1$, and so the user's payoff would be

$$\begin{aligned}
\overline{U}(p^{\delta_{\hat{q}_1}}, q') &= \varepsilon \cdot \frac{|\mathcal{Z}^+ \cap \mathcal{Z}_A|}{|\mathcal{Z}|} + (1-\varepsilon) \cdot \frac{|\mathcal{Z}^+ \cap \mathcal{Z}_A|}{|\mathcal{Z}_A|} \\
&> \varepsilon \cdot \frac{|\mathcal{Z}^+ \cap \mathcal{Z}_A|}{|\mathcal{Z}|} + \frac{|\mathcal{Z}^+|}{|\mathcal{Z}|},
\end{aligned}$$

and so playing the strategy q' guarantees the user strictly higher long-run payoff than q^{BR}.

For the forward direction, suppose without loss of generality that $\mathcal{Z}^+ \subset \mathcal{Z}_A$. Observe that conditioned on a fixed distribution of propositions $p^\mu \in \Delta(Z)$ (i.e., in the absence of any learning), q^{BR} by definition yields optimal expected payoff $\overline{U}(p^\mu, q)$. As a result, $\overline{U}(p^\mu, q^{\mathrm{BR}})$ gives an upper bound for the user's payoff when the platform's belief is μ. Now, if the platform belief is $\mu(\hat{q}_3) = 1$, then the resulting distribution is a uniform distribution over items Z, and so

$$\max_{q \in \mathcal{Q}} \overline{U}(p^{\delta_{\hat{q}_3}}, q) \leqslant \overline{U}(p^{\delta_{\hat{q}_3}}, q^{\mathrm{BR}}) = \frac{|\mathcal{Z}^+|}{|\mathcal{Z}|} = \frac{|\mathcal{Z}^+|}{|\mathcal{Z}_A| + |\mathcal{Z}_B|}.$$

Similarly, if the platform believes $\mu(\hat{q}_2) = 1$, then it will recommend uniformly with probability ε, and otherwise recommend from $\mathcal{Z}_B$, and so

$$\max_{q \in \mathcal{Q}} \overline{U}(p^{\delta_{\hat{q}_2}}, q) \leqslant \overline{U}(p^{\delta_{\hat{q}_2}}, q^{\mathrm{BR}}) = \varepsilon \cdot \frac{|\mathcal{Z}^+|}{|\mathcal{Z}|} + (1-\varepsilon) \cdot 0 = \varepsilon \cdot \frac{|\mathcal{Z}^+|}{|\mathcal{Z}|}.$$

Now, by Proposition 5.1, playing any non-degenerate strategy (i.e., q for which $\max_Z q(B = 1|Z) > 0$) will lead the platform to converge to a point mass belief. The naive strategy q^{BR} will lead to the platform having limiting belief $\hat{q}_1$, for which the user's payoff is

$$\overline{U}(p^{\delta_{\hat{q}_1}}, q^{\mathrm{BR}}) = \varepsilon \cdot \frac{|\mathcal{Z}^+|}{|\mathcal{Z}|} + (1-\varepsilon) \cdot \frac{|\mathcal{Z}^+|}{|\mathcal{Z}_A|}.$$

This is clearly greater than the other two upper bounds, and so if the user plays a non-degenerate strategy, q^{BR} is the best option. The proof concludes by noting that the degenerate strategy yields a utility of zero. □

A.3 Proof of Proposition 5.3

Proposition 5.3. *Consider the setting described in Section 5.1. The platform's payoff is as least as high when the user is strategic as when the user is naive.*

Proof. Let μ_S be the platform's long-run belief when the user is strategic, and let μ_{BR} be the platform's long-run belief when the user is naive. Note that Proposition 5.1 implies that these beliefs exist and are unique for q^{BR} and for any non-degenerate q^*. Now, if the user is incentivized to strategize, it must be that $\overline{U}(p^{\mu_{\text{BR}}}, q^{\text{BR}}) \leqslant \overline{U}(p^{\mu_S}, q^*)$. In other words,

$$\begin{aligned}
\sum_{Z\in\mathcal{Z}} p^{\mu_S}(Z)\mathbb{E}_{q^S(\cdot|Z)}[U(Z,B)] &\geqslant \sum_{Z\in\mathcal{Z}} p^{\mu_{\text{BR}}}(Z)\mathbb{E}_{q^{\text{BR}}(\cdot|Z)}[U(Z,B)] \\
\sum_{Z\in\mathcal{Z}} p^{\mu_S}(Z)q^S(B=1|Z)a(Z) &\geqslant \sum_{Z\in\mathcal{Z}} p^{\mu_{\text{BR}}}(Z)q^{\text{BR}}(B=1|Z)\cdot a(Z) \\
\sum_{Z\in\mathcal{Z}^+} p^{\mu_S}(Z)q^S(B=1|Z) - \sum_{Z\in\mathcal{Z}\setminus\mathcal{Z}^+} p^{\mu_S}(Z)q^S(B=1|Z) &\geqslant \sum_{Z\in\mathcal{Z}^+} p^{\mu_{\text{BR}}}(Z)q^{\text{BR}}(B=1|Z) \\
\sum_{Z\in\mathcal{Z}} p^{\mu_S}(Z)q^S(B=1|Z) - \sum_{Z\in\mathcal{Z}} p^{\mu_{\text{BR}}}(Z)q^{\text{BR}}(B=1|Z) &\geqslant 2\sum_{Z\in\mathcal{Z}\setminus\mathcal{Z}^+} p^{\mu_S}(Z)q^S(B=1|Z)
\end{aligned}$$

□

A.4 Proof of Proposition 5.4

Proposition 5.4. *Consider the setting described in Section 5.1, and the counterfactual algorithm p_{CF} given by (8). For any content partitioning $(\mathcal{Z}_A, \mathcal{Z}_B)$ where $|\mathcal{Z}_A|, |\mathcal{Z}_B| \geqslant 4$, there exists an affinity function $a(Z)$ (see (4)), constants $\gamma > 0$ (see (5)) and $\varepsilon > 0$ (see (6)) and a function* TOXICITY $: \mathcal{Z} \to \{\alpha, 1\}$ *such that, by applying the strategy above:*

(a) the platform correctly predicts its own utility under p, i.e., $\widehat{V}(p, \widehat{\mathcal{Q}}) = \overline{V}^*(p, \widehat{\mathcal{Q}})$;

(b) the platform thinks its payoff will decrease if it switches to algorithm p_{CF}, i.e., $\widehat{V}(p_{\text{CF}}, \widehat{\mathcal{Q}}) < \overline{V}^*(p, \widehat{\mathcal{Q}})$;

(c) in reality, the platform's payoff will increase if it switches to p_{CF}, i.e., $\overline{V}^\star(p_{\text{CF}}, \widehat{\mathcal{Q}}) > \overline{V}^\star(p, \widehat{\mathcal{Q}})$.

Proof. For convenience, assume $|\mathcal{Z}_B|$ is even (otherwise, the same proof holds but requires us to keep track of a rounding error). We will define $a(Z)$ to be $+1$ on a half of $\mathcal{Z}_B$ and -1 on the other half, and $+1$ on $3/4$ of $\mathcal{Z}_A$ (and -1 on the remaining $1/4$). By construction, defining $\mathcal{Z}^+ = \{Z \in \mathcal{Z} : a(Z) = 1\}$,

$$\frac{|\mathcal{Z}_A \cap \mathcal{Z}^+|}{|\mathcal{Z}_A|} \geqslant \frac{3}{4} > \frac{1}{2} = \frac{|\mathcal{Z}_B \cap \mathcal{Z}^+|}{|\mathcal{Z}_B|}.$$

Using an identical argument to the one used in Appendix A.2, we see that when ε is sufficiently small, a strategic user will restrict their clicks to $\mathcal{Z}^+ \cap \mathcal{Z}_1$, in order to induce a platform belief of $\hat{q}_1$. For some small $\alpha > 0$ (to be defined later), consider the toxicity function

$$\text{TOXICITY}(Z) = \begin{cases} \alpha & \text{if } Z \in \mathcal{Z}_A \cap \mathcal{Z}^+ \text{ or } \mathcal{Z}_B \setminus \mathcal{Z}^+ \\ 1 & \text{otherwise.} \end{cases}$$

In the remainder of the proof, we introduce the following notation for convenience:

$$n_1 = |\mathcal{Z}_A \cap \mathcal{Z}^+|, \qquad n_2 = |\mathcal{Z}_A \setminus \mathcal{Z}^+| \qquad n_3 = |\mathcal{Z}_B \cap \mathcal{Z}^+|, \qquad n_4 = |\mathcal{Z}_B \setminus \mathcal{Z}^+|.$$

Platform's current payoff. The platform's current payoff under the user's strategic behavior is

$$\overline{V}^*(p_0, \widehat{\mathcal{Q}}) = \varepsilon \cdot \frac{n_1}{n_1 + n_2 + n_3 + n_4} + (1 - \varepsilon) \cdot \frac{n_1}{n_1 + n_2}. \tag{17}$$

Note that the platform's predicted payoff under p_0 is given by

$$\widehat{V}(p_0, \widehat{\mathcal{Q}}) = (1 - \gamma) \left[\varepsilon \cdot \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4} + (1 - \varepsilon), \right],$$

and by setting γ appropriately we can make these two quantities equal (intuitively, since ε is small, $1 - \gamma$ roughly corresponds to the fraction of $\mathcal{Z}_A$ that the user actually engages with, i.e., $\frac{n_1}{n_1+n_2}$).

Platform's predicted payoff. We now derive the platform's predicted payoff $\widehat{V}(p_1, \widehat{\mathcal{Q}})$—the platform believes that the user's strategy is $\hat{q}_1$, and so its predicted payoff is

$$\widehat{V}(p_1, \widehat{\mathcal{Q}}) = (1 - \gamma) \left[\varepsilon \cdot \frac{\alpha n_1 + n_2}{\alpha n_1 + n_2 + n_3 + \alpha n_4} + (1 - \varepsilon) \right].$$

From this, it is straightforward to show that

$$\begin{aligned}
\widehat{V}(p_1, \widehat{\mathcal{Q}}) - \overline{V}^*(p_0, \widehat{\mathcal{Q}}) &= \widehat{V}(p_1, \widehat{\mathcal{Q}}) - \widehat{V}(p_0, \widehat{\mathcal{Q}}) \\
&= (1 - \gamma) \cdot \varepsilon \cdot \left[\frac{\alpha n_1 + n_2}{\alpha n_1 + n_2 + n_3 + \alpha n_4} - \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4} \right] \\
&< 0,
\end{aligned}$$

where in the last inequality we use the fact that:

$$\begin{aligned}
\frac{n_2}{n_1 + n_2} &\leqslant \frac{1}{4} < \frac{1}{2} = \frac{n_3}{n_3 + n_4} \\
n_2 n_3 + n_2 n_4 &< n_1 n_3 + n_2 n_3 \\
0 &> (1 - \alpha)(n_2 n_4 - n_1 n_3) \\
&= (\alpha n_1 + n_2)(n_1 + n_2 + n_3 + n_4) - (n_1 + n_2)(\alpha n_1 + n_2 + n_3 + \alpha n_4).
\end{aligned}$$

This sequence of calculations proves the result (b).

Platform's true payoff. We now consider the platform's true payoff when the user is strategic. Now, we use the fact that (a) if the user's behavior is non-degenerate, the platform's belief will converge to a single $\mu(\hat{q}_i) = 1$; (b) for a fixed belief μ, the maximum user payoff is upper bounded by the payoff attained by q^{BR}.

Thus, under the toxicity function above,

$$\begin{aligned}
\max_{q \in \mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_1}}, q) &\leqslant \overline{U}(p_1^{\delta_{\hat{q}_1}}, q^{\mathrm{BR}}) = \varepsilon \cdot \frac{\alpha n_1 + n_3}{\alpha n_1 + n_2 + n_3 + \alpha n_4} + (1 - \varepsilon) \cdot \frac{\alpha n_1}{\alpha n_1 + n_2} \\
\max_{q \in \mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_2}}, q) &\leqslant \overline{U}(p_1^{\delta_{\hat{q}_2}}, q^{\mathrm{BR}}) = \varepsilon \cdot \frac{\alpha n_1 + n_3}{\alpha n_1 + n_2 + n_3 + \alpha n_4} + (1 - \varepsilon) \cdot \frac{n_3}{n_3 + \alpha n_4} \\
\max_{q \in \mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_3}}, q) &\leqslant \overline{U}(p_1^{\delta_{\hat{q}_3}}, q^{\mathrm{BR}}) = \frac{\alpha n_1 + n_3}{\alpha n_1 + n_2 + n_3 + \alpha n_4}.
\end{aligned}$$

Observe that as $\alpha \to 0$,

$$\max_{q\in\mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_1}}, q) \to \frac{\varepsilon \cdot n_3}{n_2+n_3} \qquad \text{and} \qquad \max_{q\in\mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_3}}, q) \to \frac{n_3}{n_2+n_3}. \tag{18}$$

Also, if the user plays the strategy $q^\dagger(B=1|Z) = \mathbf{1}\{Z \in \mathcal{Z}^+ \cap \mathcal{Z}_B\}$, the platform will converge to $\mu(\hat{q}_2) = 1$ (by Proposition 5.1), and so the user's utility will be

$$\overline{U}(p_1^{\delta_{\hat{q}_2}}, q^\dagger) = \varepsilon \cdot \frac{n_3}{\alpha n_1 + n_2 + n_3 + \alpha n_4} + (1-\varepsilon)\cdot \frac{n_3}{n_3 + \alpha n_4},$$

which as $\alpha \to 0$, converges to

$$\overline{U}(p_1^{\delta_{\hat{q}_2}}, q^\dagger) \to \varepsilon \cdot \frac{n_3}{n_2+n_3} + (1-\varepsilon) > \frac{n_3}{n_2+n_3} = \lim_{\alpha\to 0}\max_{q\in\mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_3}}, q).$$

In particular, by comparing this to (18), there must exist some $\alpha > 0$ such that

$$\overline{U}(p_1^{\delta_{\hat{q}_2}}, q^\dagger) > \max_{q\in\mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_3}}, q) > \max_{q\in\mathcal{Q}} \overline{U}(p_1^{\delta_{\hat{q}_1}}, q).$$

This implies that the strategic user will induce the belief $\mu(\hat{q}_2) = 1$ when the platform plays the algorithm p_1. Note that of the user strategies that induce $\mu(\hat{q}_2) = 1$, the optimal one is clearly $q^\dagger$, as any strategy that does not set $B = 1$ for $Z \in \mathcal{Z}^+ \cap \mathcal{Z}_B$ would increase its utility by doing so, and any strategy that does set $B = 1$ for $Z \in \mathcal{Z}_B \setminus \mathcal{Z}^+$ would needlessly incur a penalty.

Thus, to conclude the proof, observe that

$$\begin{aligned}\overline{V}^*(p_1, \widehat{\mathcal{Q}}) &= \overline{V}(p_1^{\delta_{\hat{q}_2}}, q^\dagger)\\ &= \varepsilon \cdot \frac{n_3}{\alpha n_1 + n_2 + n_3 + \alpha n_4} + (1-\varepsilon)\cdot\frac{n_3}{n_3+\alpha n_4},\end{aligned}$$

which, as $\alpha \to 0$ and $\varepsilon \to 0$, converges to 1. Contrasting this to (17) makes it clear that by choosing α and ε small enough, we get that $\overline{V}^*(p_1, \widehat{\mathcal{Q}}) > \overline{V}^*(p_0, \widehat{\mathcal{Q}})$. □

A.5 Proof of Proposition 5.5

Proposition 5.5. *Consider the setting described in Section 5.1. For any partitioning $(\mathcal{Z}_A, \mathcal{Z}_B)$ of $\mathcal{Z}$ there exists an affinity function $a(Z)$ (see (4)), constants $\gamma > 0$ (see (5)) and $\varepsilon > 0$ (see (6)), and a user model $\hat{q}_4$ such that when $\hat{q}_4$ is added to hypothesis class $\widehat{\mathcal{Q}}$, the platform's payoff under strategization decreases.*

Proof. Define the affinity function $a(Z)$ to be $+1$ on half of the items in $\mathcal{Z}_A$ (rounding down if $|\mathcal{Z}_A|$ is not even), and on a single item from $\mathcal{Z}_B$.

Recall from Proposition 5.2 and its proof in Appendix A.2 that a strategic user will try to induce the belief $\mu(\hat{q}_1) = 1$ from the platform by restricting their clicks to $\mathcal{Z}_A \cap \mathcal{Z}^+$ (where recall that $\mathcal{Z}^+$ is the subset of $\mathcal{Z}$ on which $a(Z) = 1$). For some $\eta > 0$ to be defined later, define

$$\hat{q}_4(B=1|Z) = 1-\eta \quad \forall\, Z \in \mathcal{Z}$$

Now, consider a strategic user in reponse to the platform strategy $(p_0, \widehat{\mathcal{Q}} \cup \{\hat{q}_4\})$. The user is faced with a choice between three cases:

(A) Force the platform to converge to $\hat{q}_3$ or $\hat{q}_4$. These are indistinguishable from the user's perspective as in either case, the resulting proposition distribution p_0^μ will converge to a uniform distribution over $\mathcal{Z}$.

(B) Induce the platform to converge to $\mu(\hat{q}_1) = 1$, in which case the limiting proposition distribution is a mixture of the uniform distribution over $\mathcal{Z}$ (with probability ε) and the uniform distribution over $\mathcal{Z}_A$ (with probability $1 - \varepsilon$).

(C) Conversely, induce the platform to converge to $\mu(\hat{q}_2) = 1$.

Note that playing naively results in case (A), and so the user's utility in that case is both upper and lower bounded by $|\mathcal{Z}^+|/|\mathcal{Z}|$. Meanwhile, since we constructed $|\mathcal{Z}_B \cap \mathcal{Z}^+| = 1$, the user's utility in case (C) is upper bounded by $1/|\mathcal{Z}_B|$. We can thus remove case (C) from consideration.

It thus remains to bound the utility of Case (B). In the absence of $\hat{q}_4$, the user can restrict their clicks to $\mathcal{Z}_A \cap \mathcal{Z}^+$ and guarantee that the platform converges to $\hat{q}_1$ (see Appendix A.1). In the presence of $\hat{q}_4$, however, we argue that the only strategy that guarantees the platform converging to $\hat{q}_1$ will also result in low user payoff.

To show this, suppose there exists a strategy q for which the platform converges to $\hat{q}_1$. If $\eta < \gamma$,

$$\begin{aligned}
&\mathrm{KL}(q||\hat{q}_1) - \mathrm{KL}(q||\hat{q}_4) \\
&= \mathbb{E}_{Z\sim p^\mu}\left[\mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{q(B|Z)}{\hat{q}_1(B|Z)}\right) - \log\left(\frac{q(B|Z)}{\hat{q}_4(B|Z)}\right)\right]\right] \\
&= \mathbb{E}_{Z\sim p^\mu}\left[\mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{\hat{q}_4(B|Z)}{\hat{q}_1(B|Z)}\right)\right]\right] \\
&= \mathbb{E}_{Z\sim p^\mu}\left[\mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{\hat{q}_4(B|Z)}{\hat{q}_1(B|Z)}\right)\right]\middle| Z \in \mathcal{Z}_A\right] \cdot \mathbb{P}_{Z\sim p^\mu}(Z \in \mathcal{Z}_A) \\
&\qquad + \mathbb{E}_{Z\sim p^\mu}\left[\mathbb{E}_{B\sim q(\cdot|Z)}\left[\log\left(\frac{\hat{q}_4(B|Z)}{\hat{q}_1(B|Z)}\right)\right]\middle| Z \in \mathcal{Z}_B\right] \cdot \mathbb{P}_{Z\sim p^\mu}(Z \in \mathcal{Z}_B) \\
&= \mathbb{E}_{Z\sim p^\mu}\left[q(B=1|Z)\cdot\log\left(\frac{1-\eta}{1-\gamma}\right) + q(B=0|Z)\cdot\log\left(\frac{\eta}{\gamma}\right)\middle| Z \in \mathcal{Z}_A\right] \cdot \mathbb{P}_{Z\sim p^\mu}(Z \in \mathcal{Z}_A) \\
&\qquad + \mathbb{E}_{Z\sim p^\mu}\left[\log(\eta)\middle| Z \in \mathcal{Z}_B\right] \cdot \mathbb{P}_{Z\sim p^\mu}(Z \in \mathcal{Z}_B) \\
&\geqslant \left(\log(1-\eta) + \mathbb{E}_{Z\sim p^\mu}\left[q(B=0|Z)\middle| Z \in \mathcal{Z}_A\right]\cdot\log\left(\frac{\eta}{\gamma}\right)\right)\cdot\mathbb{P}(Z \in \mathcal{Z}_A) + \log(\eta)\,.
\end{aligned}$$

Thus, if

$$\mathbb{E}_{Z\sim p^\mu}\left[q(B=0|Z)\middle| Z \in \mathcal{Z}_A\right] > \frac{\left(\frac{-\log(\eta)}{\mathbb{P}(Z\in\mathcal{Z}_A)} - \log(1-\eta)\right)}{\log\left(\frac{\eta}{\gamma}\right)},$$

then $\mathrm{KL}(q||\hat{q}_1) - \mathrm{KL}(q||\hat{q}_4) > 0$, and $\hat{q}_4$ thus dominates $\hat{q}_1$ at belief μ. Now, since $\mathbb{P}(Z \in \mathcal{Z}_A)$ is lower bounded by $\varepsilon|\mathcal{Z}_A|/|\mathcal{Z}|$, we can set γ small so that the above expression evaluates to $\frac{1}{|\mathcal{Z}|}$.

Also, note that since none of the user models distinguish between different elements within the partition $\mathcal{Z}_A$, the expectation on the left is equivalent to the unconditional expectation $\mathbb{E}_{Z\sim\mathrm{Unif}(\mathcal{Z}_A)}[\cdot]$.

Putting these two observations together, we have that if

$$\frac{1}{|\mathcal{Z}_A|}\sum_{Z\in\mathcal{Z}_A} q(B=0|Z) > \frac{1}{|\mathcal{Z}|}, \tag{19}$$

then $\hat{q}_4$ strictly dominates $\hat{q}_1$ at all beliefs μ, and the platform will converge to $\mu(\hat{q}_4) = 1$. Since we have by assumption that the platform converges to $\mu(\hat{q}_1) = 1$, it must be that (19) is false, and so

$$\frac{1}{|\mathcal{Z}_A|}\sum_{Z\in\mathcal{Z}_A} q(B=1|Z) \geqslant 1 - \frac{1}{|\mathcal{Z}|}.$$

By construction ($\mathcal{Z}^+$ containing exactly half of $\mathcal{Z}_A$),

$$\frac{1}{|\mathcal{Z}_A \cap \mathcal{Z}^+|} \sum_{Z \in \mathcal{Z}_A \cap \mathcal{Z}^+} q(B = 1|Z) \geqslant 1 - \frac{2}{|\mathcal{Z}|}.$$

We now argue that with such a strategy, the user can attain no more than $\frac{2}{|\mathcal{Z}|}$ utility. Since none of the user models distinguish between elements within $\mathcal{Z}_A$, a coupling argument shows that for every element $Z : a(Z) = 1$ that the user clicks on with probability δ, there is at least one other element with $a(Z') = -1$ that the user clicks on with probability $\delta(1 - \frac{2}{|\mathcal{Z}|})$ and that is recommended with equal probability to Z.

We have thus shown that the maximum attainable payoff from case (B) is $2/|\mathcal{Z}|$, which is lower than the guaranteed payoff from case (A) of $|\mathcal{Z}^+|/|\mathcal{Z}|$. A strategic user will thus choose to play naively, which in fact *lowers* platform payoff. We have thus shown that both case (B) and case (C) result in low payoffs for the user, and so the user will choose case (A). The proof concludes by observing that platform's payoff under case (A) decreases from the original strategic user. □

B Omitted Proofs: Main Results

B.1 Proof of Proposition 6.9

Proposition 6.9. *Consider a platform strategy $(p, \widehat{\mathcal{Q}})$ and suppose that $U(B, Z)$ has a unique maximizer in $\mathcal{B}$ for all Z. For a user strategy q, let $\widetilde{V}(q)$ be the platform's worst-case limiting payoff,*

$$\widetilde{V}(q) := \min_{\mu \in \Delta(S^{\infty}(q,p,\widehat{\mathcal{Q}}))} \overline{V}(p(\cdot;\mu), q), \qquad \forall q \in \mathcal{Q}, \tag{11}$$

where S^{∞} is defined in Definition 6.7. Then, user strategization strictly improves the platform's worst-case limiting payoff if

$$\widetilde{V}\left(\arg\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(S^{\infty}(q,p,\widehat{\mathcal{Q}}))} \overline{U}(p(\cdot;\mu), q)\right) > \widetilde{V}\left(\arg\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(\widehat{\mathcal{Q}})} \overline{U}(p(\cdot;\mu), q)\right). \tag{12}$$

The same is true if the min *in* (11) *is swapped out for a* max.

Proof. Note that the left-hand side of (12) is exactly the platform's payoff under a strategic user, and so it only remains to show that the right-hand side corresponds to the platform's payoff under a naive user. First, by the min-max inequality,

$$\begin{aligned}
\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(\widehat{\mathcal{Q}})} \overline{U}(p(\cdot;\mu), q) &\leqslant \min_{\mu \in \Delta(\widehat{\mathcal{Q}})} \max_{q \in \mathcal{Q}} \overline{U}(p(\cdot;\mu), q) \\
&\leqslant \min_{\mu \in \Delta(\widehat{\mathcal{Q}})} \overline{U}(p(\cdot;\mu), q^{\mathrm{BR}}),
\end{aligned}$$

and so

$$q^{\mathrm{BR}} \in \arg\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(\widehat{\mathcal{Q}})} \overline{U}(p(\cdot;\mu), q).$$

Now, suppose $q \neq q^{\mathrm{BR}}$ satisfies

$$q \in \arg\max_{q \in \mathcal{Q}} \min_{\mu \in \Delta(\widehat{\mathcal{Q}})} \overline{U}(p(\mu;\cdot), q).$$

Since $q \neq q^{\mathrm{BR}}$ and U has a unique maximizer $B^*(Z)$ for each $Z \in \mathcal{Z}$, there must exist some $Z_0 \in \mathcal{Z}$ for which $q(\cdot|Z_0) \neq \mathbf{1}\{B^*(Z_0)\}$. Define

$$q'(\cdot|Z) = \begin{cases} q(\cdot|Z) & \text{if } Z \neq Z_0 \\ \mathbf{1}\{B^*(Z_0)\} & \text{otherwise.} \end{cases}$$

Clearly, q' attains the same payoff U as q for any $Z \neq Z_0$, and when $Z = Z_0$, q' attains better payoff than q. If there exists a μ such that $p(Z_0;\mu) > 0$ we have thus reached a contradiction, and otherwise Z_0 will never by played by the platform, and so user actions under q are indistinguishable to the platform from user actions under q^{BR}.

□

B.2 Proof of Proposition 6.13

Proposition 6.13. *Consider a given platform strategy $(p, \widehat{\mathcal{Q}})$ and platform payoff function V. Suppose that $\widehat{\mathcal{Q}}$ is an ε-net (Definition 6.12) for some sufficiently small ε, that $p(\cdot;\mu)$ has full support for all μ, and that Assumption 6.4 holds. Define ζ as the maximum gap in predicted platform payoff, i.e.,*

$$\zeta(p') = \max_{\hat{q}_1,\hat{q}_2 \in \mathcal{Q}} \overline{V}(p'(\cdot;\delta_{\hat{q}_1}),\hat{q}_1) - \overline{V}(p'(\cdot;\delta_{\hat{q}_2}),\hat{q}_2).$$

Further assume that there exists $\beta \geqslant \alpha \geqslant 0$ such that $\alpha \leqslant Var_{p(\cdot;\mu)\times q}[V(B,Z)] \leqslant \beta$ for any belief μ and user strategy q. Then, for any $\varepsilon_0, \varepsilon_1 > 0$, there exists a p_{CF} and U such that $d_{\mathcal{P}}(p, p_{CF}) \leqslant \varepsilon_0$, and and

$$\min_{\mu \in S^\infty(q,p)} \left|\widehat{V}(p_{CF},\mu) - \overline{V}^\star(p_{CF})\right| \geqslant \sqrt{\zeta(p_{CF})^2 - 4(\beta-\alpha)} - \varepsilon_1. \tag{15}$$

Proof. Now, we define $\hat{q}_1$ and $\hat{q}_2$ as the lowest and highest attainable platform payoff by a user who chooses a strategy $\hat{q} \in \widehat{\mathcal{Q}}$, i.e.,

$$\hat{q}_1 = \arg\max_{\hat{q}\in\widehat{\mathcal{Q}}} \overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}), \qquad \hat{q}_2 = \arg\min_{\hat{q}\in\widehat{\mathcal{Q}}} \overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}).$$

Note that for our bound to meaningful, we must have that

$$\gamma := \frac{\overline{V}(p(\cdot;\delta_{\hat{q}_1}),\hat{q}_1) - \overline{V}(p(\cdot;\delta_{\hat{q}_2}),\hat{q}_2)}{4} - (\beta - \alpha) > 0,$$

and so we assume that this inequality holds for the remainder of the proof. Now, recall that the payoff function V is scaled so that $0 \leqslant V(B,Z) \leqslant 1$. For some constant $c \in [0,1]$ to be set later, let

$$U(B,Z) = (V(B,Z) - c)^2,$$

so that for any belief μ and user strategy q,

$$\overline{U}(p^\mu, q, c) = \mathrm{Var}_{(B,Z)\sim p^\mu \times q}[V(B,Z)] + \left(\overline{V}(p^\mu,q) - c\right)^2,$$

where we make the dependence on the unset constant c explicit for notational convenience. In turn, the function that the strategic user aims to maximize is

$$\widetilde{U}(q,c) := \min_{\mu\in S^\infty(q,p,\widehat{\mathcal{Q}})} \overline{U}(p^\mu, q).$$

When $q = \hat{q} \in \widehat{\mathcal{Q}}$ and $p(\cdot;\mu)$ has full support for all μ, $\widetilde{U}(q,c)$ simplifies to

$$\widetilde{U}(\hat{q},c) := \overline{U}(p(\cdot;\delta_{\hat{q}}),\hat{q}).$$

We now prove a few lemmata that will be useful later in the proof. First, we show that our condition on the loss landscape of $\overline{V}$ suffices to show a similar condition for the loss landscape of $\overline{U}$:

Lemma B.1. *For any proposition distribution $r \in \Delta(\mathcal{Z})$, any fixed user strategy q, and any $\varepsilon > 0$, there exists a distribution $r' \in \Delta(\mathcal{Z})$ so that $\mathcal{W}_1(r,r') \leqslant \varepsilon$ and $\delta_{r,q}(c) := \overline{U}(r,q,c) - \overline{U}(r',q,c) \neq 0$ almost everywhere (with respect to c).*

Proof. We can use Assumption 6.4 to find a distribution $r' \in \Delta(\mathcal{Z})$ such that $d(r,r') \leqslant \varepsilon$ and $\delta := \overline{V}(r',q) - \overline{V}(r,q) \neq 0$. Then, note that $\overline{U}(r,q,c) = \overline{U}(r',q,c)$ if and only if

$$\begin{aligned} \mathrm{Var}_{(B,Z)\sim r\times q}[V(B,Z)] + (\overline{V}(r,q)-c)^2 &= \mathrm{Var}_{(B,Z)\sim r'\times q}[V(B,Z)] + (\overline{V}(r',q)-c)^2 \\ \mathrm{Var}_{(B,Z)\sim r\times q}[V(B,Z)] - \mathrm{Var}_{(B,Z)\sim r'\times q}[V(B,Z)] &= 2\delta(\overline{V}(r,q)-c) + \delta^2, \end{aligned}$$

and in particular, if and only if

$$c = \overline{V}(r,q) - \frac{1}{2\delta}\left(\mathrm{Var}_{(B,Z)\sim r\times q}[V(B,Z)] - \mathrm{Var}_{(B,Z)\sim r'\times q}[V(B,Z)] - \delta^2\right),$$

which is a measure-zero set. □

Next, we show that given a collection of proposition distributions $\{r_1, \dots r_m\}$ corresponding to a set of user models $\{\hat{q}_1, \dots, \hat{q}_m\}$ such that r_i is ε_0-close to $p(\cdot;\delta_{\hat{q}_i})$, we can construct a new algorithm p' that is $2\varepsilon_0$-close to p such that p' satisfies the regularity conditions in Assumption 6.3, and $p'(\cdot;\delta_{\hat{q}_i}) = r_i$ for all $i \in [m]$.

Lemma B.2. *Given an algorithm p satisfying Assumption 6.3, a collection of proposition distributions $\{r_1, \dots r_m\}$, and a set of user models $\{\hat{q}_1, \dots, \hat{q}_m\}$ such that*

$$\mathcal{W}_1(p(\cdot;\delta_{\hat{q}_i}), r_i) \leqslant \varepsilon_0 \;\forall\; i \in [m],$$

there exists an algorithm p' such that $d_{\mathcal{P}}(p,p') \leqslant 2\varepsilon_0$ and

$$p(\cdot;\delta_{\hat{q}_i}) = r_i \forall\; i \in [m].$$

Proof. By Assumption 6.3, we know that for all $Z \in \mathcal{Z}$, $p(Z;\mu)$ is continuous in μ around point mass beliefs. This continuity also implies continuity in Wasserstein distance. Thus, there exists a single constant δ such that for any $i \in [m]$,

$$d(\delta_{\hat{q}_i}, \mu) < \delta \implies \mathcal{W}_1(p(\cdot;\delta_{\hat{q}}), p(\cdot;\mu)) \leqslant \varepsilon_0,$$

where $d(\cdot,\cdot)$ is the same distance metric with respect to which Assumption 6.3 (ii) holds. Then, define the new algorithm

$$p'(Z;\mu) = \begin{cases} r_i & \text{if } d(\delta_{\hat{q}_i}, \mu) < \delta \\ p(Z;\mu) & \text{otherwise.} \end{cases}$$

Clearly, this algorithm satisfies $p(\cdot;\delta_{\hat{q}_i}) = r_i \forall\; i \in [m]$. Furthermore, it satisfies the continuity condition since it is actually constant in the neighborhood of each point mass belief $\delta_{\hat{q}}$. Finally,

$$\begin{aligned} d_{\mathcal{P}}(p,p') &= \sup_{\mu \in \Delta(\widehat{\mathcal{Q}})} \mathcal{W}_1(p(\cdot;\mu), p'(\cdot;\mu)) \\ &= \max_{i\in[m]} \sup_{d(\mu,\delta_{\hat{q}_i})<\delta} \mathcal{W}_1(p(\cdot;\mu), r_i) \\ &\leqslant \max_{i\in[m]} \sup_{d(\mu,\delta_{\hat{q}_i})<\delta} \mathcal{W}_1(p(\cdot;\mu), p(\cdot;\delta_{\hat{q}_i})) + \mathcal{W}_1(p(\cdot;\delta_{\hat{q}_i}), r_i) \\ &\leqslant 2\varepsilon_0, \end{aligned}$$

concluding the proof. □

We now resume the main proof. For any c, we partition the set of user models into

$$\begin{aligned}\widehat{\mathcal{Q}}_{+}(c) &= \{\hat{q} \in \widehat{\mathcal{Q}} : \overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}) > c\},\\ \widehat{\mathcal{Q}}_{-}(c) &= \{\hat{q} \in \widehat{\mathcal{Q}} : \overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}) < c\},\\ \widehat{\mathcal{Q}}_{=}(c) &= \{\hat{q} \in \widehat{\mathcal{Q}} : \overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}) = c\}.\end{aligned}$$

For any value of c, we define the functions

$$M_{+}(c) = \max_{\hat{q} \in \widehat{\mathcal{Q}}_{+}(c)} \widetilde{U}(\hat{q},c), \qquad M_{-}(c) = \max_{\hat{q} \in \widehat{\mathcal{Q}}_{-}(c)} \widetilde{U}(\hat{q},c), \qquad M(c) = \max_{\hat{q} \in \widehat{\mathcal{Q}}} \widetilde{U}(\hat{q},c),$$

then let

$$Q^{*}(c) := \arg\max_{\hat{q} \in \widehat{\mathcal{Q}}} \widetilde{U}(q,c)$$

be the set of possible strategies for a strategic user, assuming they play according to one of the user models $\hat{q} \in \widehat{\mathcal{Q}}$. Observe that for any $\hat{q} \in Q^{*}(c)$,

$$\begin{aligned}\beta + (\overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}) - c)^{2} &\geqslant \widetilde{U}(\hat{q},c)\\ &\geqslant \max\left\{\widetilde{U}(\hat{q}_{1}),\widetilde{U}(\hat{q}_{2}),\right\}\\ &\geqslant \alpha + \max\left\{\left(\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p^{\delta_{\hat{q}}},\hat{q}) - c\right)^{2}, \left(\min_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p^{\delta_{\hat{q}}},\hat{q}) - c\right)^{2}\right\}\\ &\geqslant \alpha + \frac{\left(\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p^{\delta_{\hat{q}}},\hat{q}) - \min_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p^{\delta_{\hat{q}}},\hat{q})\right)^{2}}{4}\\ (\overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}) - c)^{2} &\geqslant \frac{\left(\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p^{\delta_{\hat{q}}},\hat{q}) - \min_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p^{\delta_{\hat{q}}},\hat{q})\right)^{2}}{4} - (\beta - \alpha)\\ &= \gamma > 0. \end{aligned} \tag{20}$$

In particular, this implies that $\hat{q} \notin \widehat{\mathcal{Q}}_{=}(c)$, and thus

$$M(c) = \max\{M_{+}(c), M_{-}(c)\}.$$

Now, as c increases, the set $\widehat{\mathcal{Q}}_{-}(c)$ grows and each $\widetilde{U}(\hat{q},c)$ is non-decreasing for each $\hat{q} \in \widehat{\mathcal{Q}}_{-}(c)$, and so $M -_{(} c)$ is non-decreasing in c. Similar logic shows that $M_{+}(c)$ is non-increasing in c. Furthermore, by definition, $M_{+}(c) > M_{-}(c)$ when $c = 0$ and $M_{-}(c) > M_{+}(c)$ when $c = 1$. We thus let $c_0 = \inf\{c : M_{-}(c) \geqslant M(c)\}$, and let $c_1 = \sup\{c : M_{+}(c) \geqslant M(c)\}$.

Lemma B.3. *There exists a constant $\psi > 0$ such that the function $M_{-}(c)$ is uniformly continuous over the interval $[c_0 - \psi, 1]$, and $M_{+}(c)$ is uniformly continuous over $[0, c_1 + \psi]$*

Proof. We start by considering $M_{-}(c)$. For any ψ, the interval $[c_0 - \psi, 1]$ is compact and so by the Heine-Cantor theorem it suffices to show that $M_{-}(c)$ is pointwise continuous at each c in the interval. Let $P \subset \mathbb{R}$ be the (finite) set $\{\overline{V}(p(\cdot;\delta_{\hat{q}}),\hat{q}) : \hat{q} \in \widehat{\mathcal{Q}}\}$. Now, for any c in the interval, one of the following two cases must hold:

(A) $c \notin P$, i.e., there does not exist a $\hat{q} \in \widehat{\mathcal{Q}}$ whose corresponding platform payoff is equal to c. In this case, we can always find a δ small enough such that for all $c' \in (c - \delta, c + \delta)$, the set $\widehat{\mathcal{Q}}_{-}(c') = \widehat{\mathcal{Q}}_{-}(c)$ does not change (by the finite nature of $\widehat{\mathcal{Q}}$). We then use the continuity of $\widetilde{U}(\hat{q},c)$ in c for each $\hat{q} \in \widehat{\mathcal{Q}}$, and the fact that a maximum of continuous functions is continuous to show that M_{-} is continuous at c.

(B) $c \in P$, i.e., there exists at least one user model in $\widehat{\mathcal{Q}}$ whose corresponding payoff is c (in other words, $\widehat{\mathcal{Q}}_=(c) \neq \emptyset$). Let $\hat{q} \in \widehat{\mathcal{Q}}_=(c)$ be any such model. Define $\bar{c} = c + 2\psi$, so that $\bar{c} > c_0$. We will first show by contradiction that there exists a user model $\hat{q}' \in \widehat{\mathcal{Q}}_-(\bar{c})$ such that $\widetilde{U}(\hat{q}, \bar{c}) < \widetilde{U}(\hat{q}', \bar{c})$. In particular, if this were not the case, we would have $\widetilde{U}(\hat{q}, \bar{c}) = M_-(\bar{c})$, and since $\bar{c} > c_0$ we would have that $\widetilde{U}(\hat{q}, \bar{c}) \geqslant M(\bar{c})$, and thus by (20),

$$(\overline{V}(p(\cdot; \delta_{\hat{q}}), \hat{q}) - \bar{c})^2 = \gamma > 0.$$

By supposition, $\overline{V}(p(\cdot; \delta_{\hat{q}}), \hat{q}) = c$. Thus, by setting ψ small enough (i.e., as $\bar{c} \to c$), we reach a contradiction. We can also set ψ small enough so that $\widehat{\mathcal{Q}}_-(\bar{c}) = \widehat{\mathcal{Q}}_-(c) \cup \widehat{\mathcal{Q}}_=(c)$. In this case, the logic above applies to any $\hat{q} \in \widehat{\mathcal{Q}}_=(c)$, and so it must be that for some $\hat{q}' \in \widehat{\mathcal{Q}}_-(c)$,

$$\eta := \min_{\hat{q} \in \widehat{\mathcal{Q}}_=(c)} \left(\widetilde{U}(\hat{q}', \bar{c}) - \widetilde{U}(\hat{q}, \bar{c}) \right) > 0,$$

For the same $\hat{q}'$ (and again, any $\hat{q} \in \widehat{\mathcal{Q}}_=(c)$), and any $c' \in (c, \bar{c})$,

$$\widetilde{U}(\hat{q}', c') - \widetilde{U}(\hat{q}, c') \geqslant \eta + (\widetilde{U}(\hat{q}', \bar{c}) - \widetilde{U}(\hat{q}', c')) - (\widetilde{U}(\hat{q}, \bar{c}) - \widetilde{U}(\hat{q}, c')).$$

Observing that

$$\begin{aligned} \widetilde{U}(\hat{q}', \bar{c}) - \widetilde{U}(\hat{q}', c') &= 2(\bar{c} - c')(\overline{V} - c) + (\bar{c} - c')^2 \\ &= 4\psi(\overline{V} - c) + 4\psi^2, \end{aligned}$$

we can again set ψ small enough so that $\widetilde{U}(\hat{q}', c') > \widetilde{U}(\hat{q}, c')$. As a result, we have shown that for any $c' \in (c, \bar{c})$, the maximizer corresponding to $M_-(c')$ is not a member of $\widehat{\mathcal{Q}}_=(c)$, i.e.,

$$\arg \max_{\hat{q} \in \widehat{\mathcal{Q}}_-(c')} \widetilde{U}(\hat{q}, c') \cap \widehat{\mathcal{Q}}_=(c) = \emptyset.$$

Now, there exists a δ such that for all $c' \in [c - \delta, c + \delta]$,

$$\widehat{\mathcal{Q}}_-(c') = \begin{cases} \widehat{\mathcal{Q}}_-(c) & \text{if } c' < c \\ \widehat{\mathcal{Q}}_-(c) \cap \widehat{\mathcal{Q}}_=(c) & \text{if } c' > c, \end{cases}$$

and in both cases $\arg\max_{\hat{q} \in \widehat{\mathcal{Q}}_-(c')} \subset \widehat{\mathcal{Q}}_-(c)$, and thus we can use continuity of $\widetilde{U}(\hat{q}, c)$ in c.

The same logic implies that $M_+(c)$ is uniformly continuous on the interval $[0, c_1 + \psi]$. □

Note that by definition of c_1, and because $c_1 + \psi > c_1$, it must be that $M_+(c_1 + \psi) < M(c_1 + \psi)$ (otherwise c_1 would not be an upper bound on the set of which it is the sup), which means that $M_-(c_1 + \psi) = M(c_1 + \psi)$, which means that $c_1 + \psi \geqslant c_0$. Conversely, $c_0 - \psi \leqslant c_1$, and thus $[c_0 - \psi, c_1 + \psi]$ is a well-defined interval on which both $M_-(c)$ and $M_+(c)$ are uniformly continuous.

Consider the function $h(c) := M_+(c) - M_-(c)$ on the interval $[c_0 - \psi, c_1 + \psi]$. At the beginning of the interval, $h(c) > 0$, while at the end of the interval $h(c) < 0$. By the intermediate value theorem, there exists $c^* \in [c_0 - \psi, c_1 + \psi]$ such that $h(c^*) = 0$.

Now, $\widehat{\mathcal{Q}}^*(c^*)$ is surely not a singleton, as it must contain at least one element $\hat{q}^+ \in \widehat{\mathcal{Q}}_+(c^*)$ and one element $\hat{q}^- \in \widehat{\mathcal{Q}}_-(c^*)$. We apply the following procedure to each $\hat{q} \in \widehat{\mathcal{Q}}^*(c^*)$:

1. Use Lemma B.1 to find a distribution $r_{\hat{q}} \in \Delta(\mathcal{Z})$ such that $\mathcal{W}_1(r_{\hat{q}}, p(\cdot; \delta_{\hat{q}})) < \varepsilon_0$, and $\overline{U}(r_{\hat{q}}, \hat{q}) \neq \overline{U}(p(\cdot; \delta_{\hat{q}}), \hat{q})$. For simplicity, we assume that c^* is not one of the finite number of values of c such that we cannot apply Lemma B.1—if this is not the case, we can simply perturb c^* by some sufficiently small amount so as not to affect the calculations in the rest of the proof.
2. If $\overline{U}(r_{\hat{q}}, \hat{q}) > \overline{U}(p(\cdot; \delta_{\hat{q}}), \hat{q})$, terminate.
3. Otherwise, continue to the next $\hat{q} \in \widehat{\mathcal{Q}}^*(c^*)$.
4. If that is the last $\hat{q} \in \widehat{\mathcal{Q}}^*(c^*)$, terminate and do not perform Step 1.

At the end of the procedure, we will have some $\hat{q}^* \in \widehat{\mathcal{Q}}^*(c^*)$ such that $\hat{q}^*$ is strictly preferred by the user to any other $\hat{q} \in \widehat{\mathcal{Q}}^*(c^*)$ (and thus, to any $\hat{q} \in \widehat{\mathcal{Q}}$) under the constructed proposition distributions. In particular, we can use Lemma B.2 to construct a new algorithm p' such that

$$d_{\mathcal{P}}(p, p') < \varepsilon_0 \qquad \text{and} \qquad \arg\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{U}(p'(\cdot; \delta_{\hat{q}}), \hat{q}) = \{\hat{q}^*\}.$$

Without loss of generality, suppose that $\hat{q}^* \in \widehat{\mathcal{Q}}_-(c^*)$.

Now, let $\varepsilon_2 > 0$ be a small enough constant to ensure that $Q^*(c^* - \varepsilon_2) \subset Q_+(c^*)$ (note that $Q^*(c^* - \varepsilon_2) \subset Q_+(c^* - \varepsilon_2)$ for all $\varepsilon_2 > 0$; then, by setting ε_2 small enough we can also ensure that $Q_+(c^* - \varepsilon_2) = Q_+(c^*)$). Thus, we can set ε_2 small enough to ensure the following two conditions:

(a) $\arg\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{U}(p(\cdot; \delta_{\hat{q}}), \hat{q}) \subset \widehat{\mathcal{Q}}_+(c^*)$, and

(b) $\arg\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{U}(p'(\cdot; \delta_{\hat{q}}), \hat{q}) \subset \widehat{\mathcal{Q}}_-(c^*)$.

In particular, in combination with (20), these two conditions imply that

(a) For all $\hat{q}^*(p) \in \arg\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{U}(p(\cdot; \delta_{\hat{q}}), \hat{q})$, we have $\overline{V}(p(\cdot; \delta_{\hat{q}}), \hat{q}) \geqslant c^* + \sqrt{\gamma'}$

(b) For all $\hat{q}^*(p') \in \arg\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{U}(p'(\cdot; \delta_{\hat{q}}), \hat{q})$, we have $\overline{V}(p(\cdot; \delta_{\hat{q}}), \hat{q}) \leqslant c^* - \sqrt{\gamma'}$,

where we define

$$\gamma' := \frac{\max_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p'(\cdot; \delta_{\hat{q}}), \hat{q}) - \min_{\hat{q} \in \widehat{\mathcal{Q}}} \overline{V}(p'(\cdot; \delta_{\hat{q}}), \hat{q})}{4} - (\beta - \alpha) > 0,$$

which can be made arbitrarily close to γ by setting ε_0 small enough. Putting these two together, we get that

$$\left| \overline{V}(p(\cdot; \delta_{\hat{q}^*(p')}), \hat{q}^*(p')) - \overline{V}(p(\cdot; \delta_{\hat{q}^*(p)}), \hat{q}^*(p)) \right| \geqslant 2\sqrt{\gamma'}.$$

To conclude the proof, we show that for a sufficiently granular ε-net, the user does not gain much by deviating from the set of user models $\widehat{\mathcal{Q}}$.

Lemma B.4. *Fix any any user strategy q and any full-support algorithm p, and suppose that $\widehat{\mathcal{Q}}$ is an ε-net hypothesis class for some $\varepsilon \in (0, \frac{1}{|\mathcal{B}|})$. Then for every user model $\hat{q} \in \widehat{\mathcal{Q}}$,*

$$\hat{q} \in S^\infty(q, p, \widehat{\mathcal{Q}}) \implies \max_{Z \in \mathcal{Z}} KL(q(\cdot|Z), \hat{q}(\cdot|Z)) \leqslant \log\left(\frac{1}{1 - |\mathcal{B}| \cdot \varepsilon}\right),$$

and as a result, $\widetilde{U}(\hat{q}, c) \geqslant \widetilde{U}(q, c) - \sqrt{-\frac{1}{2}\log(1 - |\mathcal{B}|\varepsilon)}$.

Proof. Note that because $\widehat{\mathcal{Q}}$ is the Cartesian product of $\Delta_\varepsilon(\mathcal{B})$ across Z, when p has full support every element $\hat{q} \in S^\infty(q, p, \widehat{\mathcal{Q}})$ must satisfy

$$\hat{q}(\cdot|Z) \in \arg\min_{p_B \in \Delta_\varepsilon(\mathcal{B})} \mathrm{KL}(q(\cdot|Z), p_B) \;\forall\; Z \in \mathcal{Z}.$$

In particular, we can find a $\hat{q}'$ that strictly dominates any $\hat{q}$ that violates this condition by swapping its behavior at violating values of Z. Now, for $\nu = \varepsilon \cdot |\mathcal{B}|$, let q_ν be a mixture of q with the uniform distribution over $\mathcal{B}$, i.e.,

$$q_\nu(\cdot|Z) := \nu \cdot \frac{1}{|\mathcal{B}|} + (1-\nu) \cdot q(\cdot|Z).$$

By definition of the ε-net, there must be some $\hat{q}_i$ such that $\hat{q}_i(B|Z) \geqslant q_\nu(B|Z) - \varepsilon$, and in turn,

$$\mathrm{KL}(q(\cdot|Z), \hat{q}_i(\cdot|Z)) \leqslant \mathbb{E}_{B \sim q(\cdot|Z)} \left[\log \left(\frac{q(\cdot|Z)}{\nu \cdot \frac{1}{|\mathcal{B}|} + (1-\nu) \cdot q(\cdot|Z) - \varepsilon} \right) \right] = \log \left(\frac{1}{1-\nu} \right).$$

Next, observe that the function $U(B, Z) = (V(B, Z) - c)^2$ is bounded in $[0, 1]$, and so

$$\begin{aligned}
\widetilde{U}(q, c) &= \min_{\mu \in S^\infty(q,p,\widehat{\mathcal{Q}})} \overline{U}(p(\cdot;\mu), q) \\
&\leqslant \overline{U}(p(\cdot;\delta_{\hat{q}}), q) && \text{(since } \hat{q} \in S^\infty\text{)} \\
&\leqslant \overline{U}(p(\cdot;\delta_{\hat{q}}), \hat{q}) + \max_{Z \in \mathcal{Z}} \mathrm{TV}(\hat{q}(\cdot|Z), q(\cdot|Z)) && \text{(since } U \text{ is bounded in } [0,1]\text{)} \\
&\leqslant \widetilde{U}(\hat{q}, c) + \sqrt{\frac{1}{2} \log \left(\frac{1}{1-\nu} \right)},
\end{aligned}$$

where above we used the definition of total variation distance, as well as the fact that for any two probability distributions A and B, $\mathrm{TV}(A, B) < \sqrt{\frac{1}{2}\mathrm{KL}(A, B)}$ by Pinsker's inequality. □

For any $\varepsilon_1 > 0$, we can use Lemma B.4 and set

$$\varepsilon = \frac{1 - \exp\left(-\varepsilon_1^2\right)}{|\mathcal{B}|}$$

to get that $\widetilde{U}(\hat{q}, c) \geqslant \widetilde{U}(q, c) - \varepsilon_1$. Thus, if $\hat{q} \in S^\infty(q, p, \widehat{\mathcal{Q}})$ for some strategy q, then q cannot yield a significantly higher payoff than $\hat{q}$. This allows us to reduce the case of picking the optimal user strategy q to picking the optimal user model $\hat{q} \in \widehat{\mathcal{Q}}$. That is, if the user *strictly* prefers a user model $\hat{q}^*$ to any other user model, we can set ε sufficiently small so that the user strictly prefers their globally stable set to be $\{\hat{q}^*\}$, which entails playing a strategy close to $\hat{q}^*$.

Thus, the smallest possible gap between the platform's predicted payoff under p' and its true payoff under p' is given by

$$\begin{aligned}
\min_{\mu \in \Delta(\{\hat{q}^*(p)\})} \left| \widehat{V}(p', \mu) - \overline{V}^*(p') \right| &= \left| \widehat{V}(p', \delta_{\hat{q}^*(p)}) - \overline{V}(p'(\cdot;\delta_{\hat{q}^*(p')}, q^*(p'))) \right| \\
&= \left| \overline{V}(p'(\cdot;\delta_{\hat{q}^*(p)}), \hat{q}^*(p)) - \overline{V}(p'(\cdot;\delta_{\hat{q}^*(p')}, q^*(p'))) \right| \\
&\geqslant 2\sqrt{\gamma'} - \left| \overline{V}(p'(\cdot;\delta_{\hat{q}^*(p')}, \hat{q}^*(p'))) - \overline{V}(p'(\cdot;\delta_{\hat{q}^*(p')}, q^*(p'))) \right| \\
&\geqslant 2\sqrt{\gamma'} - \varepsilon_1,
\end{aligned}$$

where the last inequality follows from Lemma B.4.

□

B.3 Proof of Proposition 6.14

We prove this result using a highly oversimplified example in order to illustrate the main principle behind the proof. First, let the user's payoff function be $U(B, Z) = V(B, Z) + \lambda \cdot g(Z)$, where $V(B, Z)$ is the platform payoff and $g(Z)$ is a function to be specified later (along with the scalar λ). Even without specifying $g(Z)$, it is clear that for any $Z \in \mathcal{Z}$, $q^{\mathrm{BR}}(\cdot|Z) = \arg\max_{B \in \mathcal{B}} V(B, Z)$.

Now, find the following two user models $\hat{q}_1$ and $\hat{q}_2$:

$$\hat{q}_1 = \arg\min_{\hat{q} \in \widehat{\mathcal{Q}}} \mathrm{KL}\left[p(\cdot; \delta_{\hat{q}}) \times q^{\mathrm{BR}}, p(\cdot; \delta_{\hat{q}}) \times \hat{q}\right],$$
$$\hat{q}_2 = \arg\max_{\hat{q} \in \widehat{\mathcal{Q}}} \mathcal{W}\left[p(\cdot; \delta_{\hat{q}_1}), p(\cdot; \delta_{\hat{q}})\right].$$

(In the literature, $\hat{q}_1$ is referred to as a Berk-Nash equilibrium [EP16; FII20].) Consider the distributions $p(\cdot; \delta_{\hat{q}_1})$ and $p(\cdot; \delta_{\hat{q}_2})$, and define

$$g(\cdot) := \arg \max_{\|f\|_L \leqslant 1} \mathbb{E}_{Z \sim p(\cdot; \delta_{\hat{q}_2})}[f(Z)] - \mathbb{E}_{Z \sim p(\cdot; \delta_{\hat{q}_1})}[f(Z)],$$

where $\|f\|_L$ represents the Lipschitz constant of the function $f : \mathcal{Z} \to \mathbb{R}$.

By construction of g, we have that for any user strategy q,

$$\begin{aligned}\overline{U}(p(\cdot; \delta_{\hat{q}_2}), q) - \overline{U}(p(\cdot; \delta_{\hat{q}_1}), q) &\geqslant \lambda \cdot \mathcal{W}(p(\cdot; \delta_{\hat{q}_1}), p(\cdot; \delta_{\hat{q}_2})) - 1 \\ &\geqslant \frac{\lambda}{2} \max_{\hat{q}, \hat{q}' \in \widehat{\mathcal{Q}}} \mathcal{W}(p(\cdot; \delta_{\hat{q}}), p(\cdot; \delta_{\hat{q}'})) - 1 \\ &> 0,\end{aligned}$$

as long as $\lambda > \frac{2}{R}$.

Now, let $\widehat{\mathcal{Q}}_1 = \{\hat{q}_2\}$ and $\widehat{\mathcal{Q}}_2 = \{\hat{q}_2, \hat{q}_1\}$. Under $\widehat{\mathcal{Q}}_1$, the platform will trivally converge to $\hat{q}_2$ regardless of the user's behavior, and so the user is incentivized to play according to q^{BR}.

When $\hat{q}_1$ is added to the set of user models, the naive user strategy q^{BR} will lead to the user model $\hat{q}_1$ never being eliminated from the globally stable set (due to its status as a Berk-Nash equilibrium for q^{BR}). By construction of the user payoff function, this is always suboptimal for the user, since $\delta_{\hat{q}_1}$ will be in $\Delta(S^\infty(q, p, \widehat{\mathcal{Q}}))$, and so they will be incentivized to switch to a strategy that ensures only $\hat{q}_2$ is in the globally stable set.

Since, by assumption, $V(B, Z)$ has a unique maximizer for each $Z \in \mathcal{Z}$, the strategic user's new strategy must result in strictly lower platform payoff, concluding the proof of the theorem.

Remark B.5. *Note that while the example presented in this proof is rather contrived, the principle behind it is actually quite general. The principle is that the strategic user wants to induce a specific proposition distribution from the platform, but does not want to stray too far from their best-response behavior. Thus, they purposefully pick a strategy that the platform misinterprets (in this case, that strategy is just q^{BR}, but it could be any strategy q_1^*). When the platform gets "better" at capturing their behavior by adding a user model that is close to q_1^*, the user is forced to move even further away from their best-response behavior, making things worse for the platform.*

B.4 Proof of Proposition 6.15

Proposition 6.15. *Suppose that the platform's strategy $(p, \widehat{\mathcal{Q}})$ is such that the hypothesis class $\widehat{\mathcal{Q}}$ is an ε-net (Definition 6.12). Let p_{CF} be a counterfactual algorithm that is well-behaved (Assumption 6.5); then,*

$$\max_{\mu \in \Delta(S^\infty(q^{BR}, p))} \left| \widehat{V}(p_{CF}, \mu) - \overline{V}_{BR}(p_{CF}) \right| \leqslant \sqrt{\varepsilon} \left((2L_{\mathcal{P}} + 1) \sqrt{|\mathcal{B}|} \right). \tag{16}$$

As a result, by using a sufficiently fine ε-net hypothesis class $\widehat{\mathcal{Q}}$, the platform can estimate its payoff under counterfactual algorithms up to arbitrary precision.

Proof. We first restate the two quantities being compared:

$$\begin{aligned} \widehat{V}(p_{\text{CF}}, \mu) &= \mathbb{E}_{\hat{q} \sim \mu} \left[\overline{V}(p_{\text{CF}}(\cdot; \mu), \hat{q}) \right]. \\ \overline{V}_{\text{BR}}(p_{\text{CF}}) &= \min_{\mu \in \Delta(S^\infty(q^{\text{BR}}, p_{\text{CF}}))} \overline{V} \left(p_{\text{CF}}(\cdot; \mu), q^{\text{BR}} \right). \end{aligned}$$

Using the triangle inequality,

$$\begin{aligned} \max_{\mu \in \Delta(S^\infty(q^{\text{BR}}, p))} \left| \widehat{V}(p_{\text{CF}}, \mu) - \overline{V}_{\text{BR}}(p_{\text{CF}}) \right| &\leqslant \max_{\mu_1, \mu_2 \in \Delta(S^\infty(q^{\text{BR}}, p))} \left| \mathbb{E}_{\hat{q} \sim \mu_1} \left[\overline{V}(p_{\text{CF}}(\cdot; \mu_1), \hat{q}) \right] - \overline{V}\left(p_{\text{CF}}(\cdot; \mu_2), q^{\text{BR}} \right) \right| \\ &\leqslant \max_{\mu_1, \mu_2 \in \Delta(S^\infty(q^{\text{BR}}, p))} \left| \mathbb{E}_{\hat{q} \sim \mu_1} \left[\overline{V}(p_{\text{CF}}(\cdot; \mu_1), \hat{q}) - \overline{V}(p_{\text{CF}}(\cdot; \mu_2), \hat{q}) \right] \right| \\ &\qquad + \left| \mathbb{E}_{\hat{q} \sim \mu_1} \left[\overline{V}(p_{\text{CF}}(\cdot; \mu_2), \hat{q}) - \overline{V}\left(p_{\text{CF}}(\cdot; \mu_2), q^{\text{BR}} \right) \right] \right| \\ &\leqslant \max_{\mu_1, \mu_2 \in \Delta(S^\infty(q^{\text{BR}}, p))} \left(\left| \mathbb{E}_{\hat{q} \sim \mu_1} \left[\overline{V}(p_{\text{CF}}(\cdot; \mu_1), \hat{q}) - \overline{V}(p_{\text{CF}}(\cdot; \mu_2), \hat{q}) \right] \right| \right. \\ &\qquad \left. + \max_{\hat{q} \in S^\infty(q^{\text{BR}}, p)} \left| \overline{V}(p_{\text{CF}}(\cdot; \mu_2), \hat{q}) - \overline{V}\left(p_{\text{CF}}(\cdot; \mu_2), q^{\text{BR}} \right) \right| \right) \end{aligned}$$

Note that the second term above is bounded by the maximum total variation distance between $\hat{q}$ and q^{BR}, which we can bound using Lemma B.4, restated below:

Lemma B.4. *Fix any any user strategy q and any full-support algorithm p, and suppose that $\widehat{\mathcal{Q}}$ is an ε-net hypothesis class for some $\varepsilon \in (0, \frac{1}{|\mathcal{B}|})$. Then for every user model $\hat{q} \in \widehat{\mathcal{Q}}$,*

$$\hat{q} \in S^\infty(q, p, \widehat{\mathcal{Q}}) \implies \max_{Z \in \mathcal{Z}} KL(q(\cdot|Z), \hat{q}(\cdot|Z)) \leqslant \log\left(\frac{1}{1 - |\mathcal{B}| \cdot \varepsilon} \right),$$

and as a result, $\widetilde{U}(\hat{q}, c) \geqslant \widetilde{U}(q, c) - \sqrt{-\frac{1}{2}\log(1 - |\mathcal{B}|\varepsilon)}$.

We can then bound the first term using the well-behavedness condition which implies that

$$\begin{aligned} d_{\mathcal{P}}(p_{\text{CF}}(\cdot; \mu_1), p_{\text{CF}}(\cdot; \mu_2)) &\leqslant L_{\mathcal{P}} \cdot \mathbb{E}_{\hat{q}_1 \sim \mu_1, \hat{q}_2 \sim \mu_2} \left[\max_{Z \in \mathcal{Z}} \text{TV}(\hat{q}_1(\cdot|Z), \hat{q}_2(\cdot|Z)) \right] \\ &\leqslant L_{\mathcal{P}} \cdot \mathbb{E}_{\hat{q}_1 \sim \mu_1, \hat{q}_2 \sim \mu_2} \left[\max_{Z \in \mathcal{Z}} \text{TV}(\hat{q}_1(\cdot|Z), q^{\text{BR}}(\cdot|Z)) + \text{TV}(\hat{q}_2(\cdot|Z), q^{\text{BR}}(\cdot|Z)) \right], \end{aligned}$$

which we can again bound using Lemma B.4. Thus,

$$\max_{\mu \in \Delta(S^\infty(q^{\text{BR}}, p))} \left| \widehat{V}(p_{\text{CF}}, \mu) - \overline{V}_{\text{BR}}(p_{\text{CF}}) \right| \leqslant (2 \cdot L_{\mathcal{P}} + 1) \sqrt{\frac{1}{2} \log\left(\frac{1}{1 - |\mathcal{B}| \cdot \varepsilon} \right)} \leqslant (2L_{\mathcal{P}} + 1) \cdot \sqrt{|\mathcal{B}| \cdot \varepsilon},$$

where the last inequality follows from $-\log(1 - x) \leqslant 2x$ for all $x \in [0, \frac{1}{2}]$. □