Improved Bayes Risk Can Yield Reduced Social Welfare Under Competition

Meena Jagadeesan, Michael I. Jordan, Jacob Steinhardt*, and Nika Haghtalab*

University of California, Berkeley

Abstract

As the scale of machine learning models increases, trends such as scaling laws anticipate consistent downstream improvements in predictive accuracy. However, these trends take the perspective of a single model-provider in isolation, while in reality providers often compete with each other for users. In this work, we demonstrate that competition can fundamentally alter the behavior of these scaling trends, even causing overall predictive accuracy across users to be non-monotonic or decreasing with scale. We define a model of competition for classification tasks, and use data representations as a lens for studying the impact of increases in scale. We find many settings where improving data representation quality (as measured by Bayes risk) decreases the overall predictive accuracy across users (i.e., social welfare) for a marketplace of competing model-providers. Our examples range from closed-form formulas in simple settings to simulations with pretrained representations on CIFAR-10. At a conceptual level, our work suggests that favorable scaling trends for individual model-providers need not translate to downstream improvements in social welfare in marketplaces with multiple model providers.

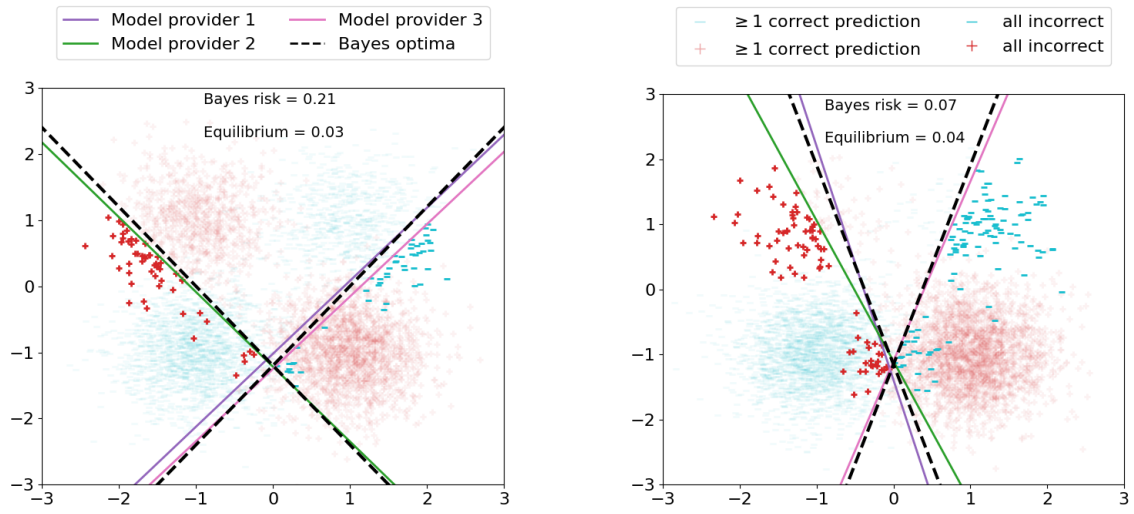
1 Introduction

Scaling trends in machine learning suggest that increasing the scale of a system consistently improves predictive accuracy. For example, scaling laws illustrate that increasing the number of model parameters [KMH+20; SK20; BDK+21] and amount of data [HBM+22] can reliably improve model performance, leading to better representations and thus better predictions for downstream tasks [HKH+21].

However, these scaling laws typically take the perspective of a single model-provider in isolation, when in reality, model-providers often compete with each other for users. For example, in digital marketplaces, multiple online platforms may provide similar services (e.g., Google search vs. Bing, Spotify vs. Pandora, Apple Maps vs. Google) and thus compete for users on the basis of prediction quality. A distinguishing feature of competing platforms is that users can switch between platforms and select a platform that offers them the highest predictive accuracy for their specific requests. This breaks the direct connection between predictive accuracy of a single platform in isolation and social welfare across competing platforms, and raises the question: *what happens to scaling laws when model-providers compete with each other?*

We show that the typical intuition about scaling laws can fundamentally break down under competition. Surprisingly, even monotonicity can be violated: increasing scale can *decrease* the overall predictive accuracy (social welfare) for users. More specifically, we study increases to scale through the lens of data representations (i.e., learned features), motivated by how increasing scale

*Equal contribution



(a) (b)

Figure 1: Comparison of equilibrium loss on two data distributions, one with high Bayes risk (left) and one with lower Bayes risk (right). Each plot shows the linear predictors chosen at equilibrium under competition between three model-providers (solid lines), along with two approximately Bayes-optimal predictors (dashed lines). The equilibrium social loss is lower in the left plot than the right plot, even though the Bayes risk is much higher. The intuition is that approximate Bayes optima disagree on more data points in the left plot than in the right plot; thus, users have a greater likelihood of at least one predictor offering them a correct prediction, which increases the overall predictive accuracy for users (i.e., the social welfare).

generally improves representation quality [BCV13].[1] We exhibit several multi-class classification tasks where better data representations (as measured by Bayes risk) *decrease* the overall predictive accuracy (social welfare) for users, when varying data representations along several different axes.

The basic intuition for this non-monotonicity is illustrated in Figure 1. When data representations are low quality, any predictor will be incorrect on a large fraction of users, and near-optimal predictors may disagree on large subpopulations of users. Model providers are thus incentivized to choose complementary predictors that cater to different subpopulations (market segments), thus improving the overall predictive accuracy for users. In contrast, when representations are high quality, each optimal predictor is incorrect on only a small fraction of users, and near-optimal predictors likely agree with each other on most data points. As a result, model-providers are incentivized to select similar predictors, which decreases the overall predictive accuracy for users.

To study when non-monotonicity can occur, we first focus on a stylized setup that permits closed-form calculations of the social welfare at equilibrium (Section 3). Using this characterization, in three concrete binary classification setups, we show that the equilibrium social welfare can be non-monotonic in Bayes risk. In particular, we vary representations along three axes—the per-representation Bayes risks, the noise level of representations, and the dimension of the data

[1] We are motivated by emerging marketplaces where different model-providers utilize the same pretrained model, but *finetune* the model in different ways. To simplify this complex training process, we conceptualize pretraining as *learning data representations (e.g., features)* and fine-tuning as *learning a predictor from these representations*. In this formalization, increasing the scale of the pretrained model (e.g., by increasing the number of parameters or the amount of data) leads to improvements in data representations accessible to the model-providers during "fine-tuning".

representations—and exhibit non-monotonicity in each case (Figure 2).

Going beyond the stylized setup of Section 3, in Section 4 we consider linear function classes and demonstrate empirically that the social welfare can be non-monotonic in the data representation quality. We consider binary and 10-class image classification tasks on CIFAR-10 where data representations are obtained from the last-layer representations of AlexNet, VGG16, ResNet18, ResNet34, and ResNet50, pretrained on ImageNet. Better representations (as measured by Bayes risk) can again perform worse under competition (Figures 4 and 5). We also consider synthetic data where we can vary representation quality more systematically, again finding ubiquitous non-monotonicities.

Altogether, our results demonstrate that the classical setting of a single model-provider can be a poor proxy for understanding multiple competing model-providers. This suggest that caution is needed when inferring that increased social welfare necessarily follows from the continuing trend towards improvements in predictive accuracy in machine learning models. Machine learning researchers and regulators should evaluate methods in environments with competing model-providers in order to reasonably assess the implications of raw performance improvements for social welfare.

1.1 Related work

Our work connects to research threads on the *welfare implications of algorithmic decisions* and *competition between data-driven platforms*.

Welfare implications of algorithmic decisions. Recent work investigates *algorithmic monoculture* [KR21; BCK+22], a setting in which multiple model-providers use the same predictor. In these works, monoculture is intrinsic to the decision-making pipeline: model-providers are given access to a shared algorithmic ranking [KR21] or shared components in the training pipeline [BCK+22]. In contrast, in our work, monoculture may arise endogenously from competition, as a result of scaling trends. Model-providers are always given access to the same function classes and data, but whether or not monoculture arises depends on the quality of data representations and its impact on the incentives of model-providers. Our work thus offers a new perspective on algorithmic monoculture, suggesting that it may arise naturally in competitive settings as a side effect of improvements in data representation quality.

More broadly, researchers have identified several sources of mismatch between predictive accuracy and downstream welfare metrics. This includes *narrowing* of a classifier under repeated interactions with users [HSN+18], *preference shaping* of users induced by a recommendation algorithm [CDR+22; DM22; CHR+22], *strategic adaptation* by users under a classifier [BKS12; HMP+16], and the *long-term impact of algorithmic decisions* [LDR+18; LWH+20].

Competition between data-driven platforms. Our work is also related to the literature on competing predictors. The model in our paper shares similarities with the work of Ben-Porat and Tennenholtz [BT17; BT19], who studied equilibria between competing predictors. Ben-Porat and Tennenholtz [BT17; BT19] show that empirical risk minimization is not an optimal strategy for a model-provider under competition and design algorithms that compute the best-responses; in contrast, our focus is on the equilibrium social welfare and how it changes with data representation quality. The specifics of our model also slightly differ from the specifics of Ben-Porat and Tennenholtz [BT17; BT19]. In their model, each user has an accuracy target that they wish to achieve and randomly chooses between model-providers that meet that accuracy target; in contrast, in our model, each user noisily chooses the model-provider that minimizes their loss and model-providers can have asymmetric market reputations.

Our work also relates to *bias-variance games* [FGH+19] between competing model-providers. However, Feng et al. [FGH+19] focus on the the equilibrium strategies for the model-provider, but do not consider equilibrium social welfare for users; in contrast, our work focuses on the equilibrium social welfare. The model of Feng et al. [FGH+19] also differs from the model in our work. In Feng et al. [FGH+19], a model-provider action is modeled as choosing an error *distribution* for each user, where the randomness in the error is intended to capture randomness in the training data samples and in the predictor; moreover, the action set includes error distributions with a range of different variances. In contrast, in our population-level setup with deterministic predictors, the error distribution for every user is always a point mass (variance 0). Thus, the equilibrium characterization of Feng et al. [FGH+19] does not translate to our model. The specifics of the model-provider utility in the work of Feng et al. [FGH+19] differs slightly from our model as well.

Other aspects studied in this research thread include competition between model-providers using *out-of-box* learning algorithms that do not directly optimize for market share [GZK+21; KGZ22; DCR+22], competition between model-providers selecting *regularization parameters* that tune model complexity [IK22], competition between *bandit algorithms* where data directly comes from users [AMS+20; JJH22], and competition between *algorithms dueling* for a user [IKL+11]. Our work also relates to *classical economic models of product differentiation* such as Hotelling's model [Hot81; dGT79] (see Anderson, de Palma, and Thisse [AdPT92] for a textbook treatment), as well as the emerging area of *platform competition* [see, e.g., JS21; CP21].

2 Model

We focus on a multi-class classification setup with input space $X \subseteq \mathbb{R}^d$ and output space $Y = \{0, 1, 2, \ldots, K-1\}$. Each user has an input x and a corresponding true output y, drawn from a distribution $\mathcal{D}$ over $X \times Y$. Model providers choose predictors f from some model family $\mathcal{F} \subseteq (\Delta(Y))^X$ where $\Delta(Y)$ is the set of distributions over Y. A user's loss given predictor f is $\ell(f(x), y) = \mathbb{P}[y \neq f(x)]$. In Section 3, we take $\mathcal{F} = \{0, 1, 2, \ldots, K-1\}^X$ to be all deterministic functions mapping inputs to classes, while in Section 4 we consider linear predictors of the form $f(x) = \text{softmax}(Wx + b)$.

We study competition between $m \geq 2$ model-providers for users, building on the model of Ben-Porat and Tennenholtz [BT17; BT19]. We index the model-providers by $[m] := \{1, 2, \ldots, m\}$, and let f_j denote the predictor chosen by model provider j. After the model-providers choose predictors $f_1, \ldots, f_m$, each user then chooses one of the m model-providers to link to, based on prediction accuracy. Model-providers aim to optimize the number of users that they win. (We note that this model is stylized and will make several simplifying assumptions; we defer a detailed discussion of the implications of these assumptions to Section 5.)

User decisions. Users noisily pick the model-provider offering the best predictions for them. That is, a user with representation x and true label y chooses a model-provider $j^*(x, y)$ such that the loss $|y - f_{j^*(x,y)}(x)|$ is the smallest across all model-providers $j \in [m]$, subject to noise in user decisions. More formally, we model user noise with the logit model [Tra02], also known as the Boltzmann rationality model:

$$\mathbb{P}[j^*(x, y) = j] = \frac{e^{-\ell(f_j(x), y)/c}}{\sum_{j'=1}^{m} e^{-\ell(f_{j'}(x), y)/c}}, \tag{1}$$

where $c > 0$ denotes a noise parameter. We extend this model to account for uneven market reputations across decisions in Section 3.5.

Model provider incentives. A model-provider's utility is captured by the market share that they win. That is, model-provider j's utility is

$$u(f_j; \mathbf{f}_{-j}) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathbb{P}[j^*(x,y) = j]\right],$$

where $\mathbf{f}_{-j}$ denotes the predictors chosen by the other model-providers and where the expectation is over (x, y) drawn from $\mathcal{D}$. Since the market shares always sum to one, this is a constant-sum game.

Each model-provider chooses a *best response* to the predictors of other model-providers. That is, model-provider j chooses a predictor f_j^* such that

$$f_j^* \in \arg\max_{f_j \in \mathcal{F}} u(f_j; \mathbf{f}_{-j}).$$

The best-response captures that model-providers optimize for market share. In practice, model-providers may do so via A/B testing to steer towards predictors that maximize profit, or by actively collecting data on market segments where competitors are performing poorly.

We study market outcomes $\mathbf{f} = (f_1^*, f_2^*, \ldots, f_m^*)$ that form a Nash equilibrium. Recall that $(f_1^*, f_2^*, \ldots, f_m^*)$ is a *pure strategy Nash equilibrium* if for every $j \in [m]$, model-provider j's predictor is a best-response to $\mathbf{f}_{-j}^*$: that is, $f_j^* \in \arg\max_{f_j \in \mathcal{F}} u(f_j; \mathbf{f}_{-j}^*)$. In well-behaved instances, pure-strategy equilibria exist (see theoretical results in Section 3 and simulation results in Section 4). However, for our results in Section 3.5, we must turn to mixed strategy equilibria where model-providers instead choose distributions μ_j over $\mathcal{F}$.

Quality of market outcome for users. We are interested in studying the quality of a market outcome $\mathbf{f} = (f_1, f_2, \ldots, f_m)$ in terms of user utility. The quality of $\mathbf{f}$ is determined by the overall *social loss* that it induces on the user population, after users choose between model-providers:

$$\texttt{SL}(f_1, \ldots f_m) := \mathbb{E}[\ell(f_{j^*(x,y)}(x), y)]. \tag{2}$$

When $f_1^*, \ldots, f_m^*$ is a Nash equilibrium, we refer to $\texttt{SL}(f_1^*, \ldots f_m^*)$ as the *equilibrium social loss.*

Our goal is to study how the equilibrium social loss changes when the representation quality (i.e., the quality of the input representations X) improves. We formalize representation quality as the minimum risk $\texttt{OPT}_{\text{single}}$ that a single model-provider could have achieved on the distribution $\mathcal{D}$ with the model family $\mathcal{F}$. This means that $\texttt{OPT}_{\text{single}}$ is equal to the Bayes risk:

$$\texttt{OPT}_{\text{single}} := \min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(f(x), y)\right].$$

In the following sections, we show that the equilibrium social loss $\texttt{SL}(f_1^*, \ldots f_m^*)$ can be non-monotonic in the representation quality (as measured by $\texttt{OPT}_{\text{single}}$), when representations are varied along a variety of axes.

3 Non-monotonicity of Equilibrium Social Loss in a Stylized Setup

To understand when non-monotonicity can occur, we first consider a stylized setup (described below) that permits closed-form calculations of the social loss. We first show a simple mathematical example that illustrates non-monotonicity (Section 3.1). We characterize the equilibrium social loss in this setup for binary classification (Section 3.2), and apply this characterization to three concrete setups that vary representation quality along different axes (Section 3.3): we show that

the equilibrium social loss can be non-monotonic in Bayes risk in all of these setups (Figures 2b-2c). Finally, we extend our theoretical characterization from Section 3.2 to setups with more than 2 classes (Section 3.4), and we extend our model and results to model-providers with unequal market reputations (Section 3.5).

Specification of stylized setup. Assume the input space X is finite and let $\mathcal{F} = \mathcal{F}_{\text{all}}^{\text{multi-class}}$ contain all deterministic functions from X to $\{0, 1, \ldots, K-1\}$. For simplicity, we also assume that users make noiseless decisions (i.e., we take $c \to 0$), so a user's choice of model-provider $j^*(x, y)$ is specified as follows:

$$\mathbb{P}[j^*(x,y) = j] = \begin{cases} 0 & \text{if } j \notin \arg\min_{j' \in [m]} \mathbb{1}[y \neq f_{j'}(x)] \\ \frac{1}{\left|\arg\min_{j' \in [m]} \mathbb{1}[y \neq f_{j'}(x)]\right|} & \text{if } j \in \arg\min_{j' \in [m]} \mathbb{1}[y \neq f_{j'}(x)]. \end{cases} \tag{3}$$

In other words, users pick the model-provider with minimum loss, choosing randomly in case of ties. We show that pure strategy equilibria are guaranteed to exist in this setup.

Proposition 1. *Let X be a finite set of representations, let there be $K \geq 2$ classes, let $\mathcal{F} = \mathcal{F}_{all}^{multi\text{-}class}$, and let $\mathcal{D}$ be the distribution over (X, Y). Suppose that user decisions are noiseless (i.e., user decisions are given by (3)). For any $m \geq 2$, there exists a pure strategy equilibrium.*

3.1 Simple mathematical example of non-monotonicity

We show a simple example where improving data representation quality (i.e. Bayes risk) reduces the equilibrium social welfare. Consider a distribution over binary labels given by $\mathbb{P}[Y = 1] = 0.6$ and $\mathbb{P}[Y = 0] = 0.4$, and suppose that there are $m = 3$ model-providers. We consider two different sets of representations X_1 and X_2, which give rise to two different distributions $\mathcal{D}_1$ over $X_1 \times Y$ and $\mathcal{D}_2$ over $X_2 \times Y$ satisfying $\mathbb{P}[Y = 1] = 0.6$ and $\mathbb{P}[Y = 0] = 0.4$.

Suppose that $X_1 = \{x_0\}$ consists of the trivial representation which provides no information about users. The distribution $\mathcal{D}_1$ is specified by $\mathbb{P}_{\mathcal{D}_1}[Y = 1 \mid X_1 = x_0] = 0.6$ and $\mathbb{P}_{\mathcal{D}_1}[Y = 0 \mid X_1 = x_0] = 0.4$. In this case, the Bayes risk is 0.4. Moreover, it is not difficult to see that $f_1^*(x_0) = f_2^*(x_0) = 1$ and $f_3^*(x_0) = 0$ is an equilibrium. (The reason that $f_1(x_0) = f_2(x_0) = f_3(x_0) = 1$ is not an equilibrium is that model provider 3 would deviate to $f_3^*(x_0) = 0$ and *increase* their utility from 1/3 to 0.4.) Since the model-providers collectively offer both labels for the representation x_0, each user has the option to choose either label, so the equilibrium social loss $\texttt{SL}(f_1^*, f_2^*, f_3^*) = 0$.

Next, suppose that $X_2 = \{x_1, x_2\}$ consists of binary representations that provide some non-trivial information about users. In particular, the distribution $\mathcal{D}_2$ is specified by equally likely representations $\mathbb{P}_{\mathcal{D}_2}[X_2 = x_1] = \mathbb{P}_{\mathcal{D}_2}[X_2 = x_2] = 0.5$. The conditional distribution $Y \mid X_2$ is specified by $\mathbb{P}_{\mathcal{D}_2}[Y = 1 \mid X_2 = x_1] = 0.4$, $\mathbb{P}_{\mathcal{D}_2}[Y = 0 \mid X_2 = x_1] = 0.6$, $\mathbb{P}_{\mathcal{D}_2}[Y = 1 \mid X_2 = x_2] = 0.8$, and $\mathbb{P}_{\mathcal{D}_2}[Y = 0 \mid X_2 = x_2] = 0.2$. In this case, the Bayes risk goes down to 0.3. Moreover, it is not difficult to see that $f_1^*(x_1) = f_2^*(x_1) = 0$, $f_3^*(x_1) = 1$, and $f_1^*(x_2) = f_2^*(x_2) = f_3^*(x_2) = 1$ is an equilibrium. (Intuitively, the reason that $f_1^*(x_2) = f_2^*(x_2) = f_3^*(x_2) = 1$ occurs at equilibrium in this setup is that no model provider $i \in [m] = \{1, 2, 3\}$ wants to deviate to $f_i(x_2) = 0$, since this would *decrease* their utility on $X_2 = x_2$ from 1/3 to 0.2.) Since users with representation x_2 no longer have the option to choose the label of 0, the equilibrium social loss is $\texttt{SL}(f_1^*, f_2^*, f_3^*) = 0.1$.

As a result, even though the Bayes risk is lower for representations in the second setup than for the representations in the first setup, the equilibrium social loss is higher. This instantation thus provides a simple mathematical example where non-monotonicity occurs. In the remaining sections, we consider more general setups that elucidate what factors drive non-monotonicity.

3.2 Characterization of the equilibrium social loss for binary classification

To generalize the above example, we analyze general instantations of the stylized setup, focusing first on binary classification. Let $\mathcal{F}_{\text{all}}^{\text{binary}}$ denote the function class $\mathcal{F}_{\text{all}}^{\text{multi-class}}$ in the special case of $K = 2$ classes. Since $\mathcal{F}_{\text{all}}^{\text{binary}}$ lets model-providers make independent predictions about each representation x, the only source of error is noise in individual data points. To capture this, we define the *per-representation Bayes risk* $\alpha(x)$ to be:

$$\alpha(x) := \min(\mathbb{P}(y = 1 \mid x), \mathbb{P}(y = 0 \mid x)). \tag{4}$$

The value $\alpha(x)$ measures how random the label y is for a given representation x. As a result, $\alpha(x)$ is the minimum error that a model-provider can hope to achieve on the given representation x. Increasing $\alpha(x)$ increases the Bayes risk $\texttt{OPT}_{\text{single}}$: in particular, $\texttt{OPT}_{\text{single}}$ is equal to the average value $\mathbb{E}[\alpha(x)]$ across the population. The equilibrium social loss, however, depends on other aspects of $\alpha(x)$.

We characterize the equilibrium social loss in terms of the per-representation Bayes risks in the following proposition. Our characterization focuses on pure-strategy equilibria, which are guaranteed to exist in this setup (see Proposition 1).

Proposition 2. *Let X be a finite set, let $K = 2$, and let $\mathcal{F} = \mathcal{F}_{all}^{binary}$. Suppose that user decisions are noiseless (i.e., user decisions are given by (3)). Suppose also that $\alpha(x) \neq 1/m$ for all $x \in X$.*[2] *At any pure strategy Nash equilibrium $f_1^*, \ldots, f_m^*$, the social loss $\texttt{SL}(f_1^*, \ldots, f_m^*)$ is equal to:*

$$\texttt{SL}(f_1^*, \ldots, f_m^*) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\alpha(x) \cdot \mathbb{1}[\alpha(x) < 1/m]] \,. \tag{5}$$

The primary driver of Proposition 2 is that as the per-representation Bayes risk $\alpha(x)$ decreases, the equilibrium predictions for x go from *heterogeneous* (different model-providers offer different predictions for x) to *homogenous* (all model-providers offer the same prediction for x). In particular, if $\alpha(x)$ is below $1/m$, then all model-providers choose the Bayes optimal label $y^* = \arg\max_{y'} \mathbb{P}[y' \mid x]$, so predictions are homogeneous; on the other hand, if $\alpha(x)$ is above $1/m$, then at least one model-provider will choose $1 - y^*$, so predictions are heterogeneous. When predictions are heterogeneous, each user is offered perfect predictive accuracy by some model-provider, which results in zero social loss. On the other hand, if predictions are homogeneous and all model-providers choose the Bayes optimal label, the social loss on x is the per-representation Bayes risk $\alpha(x)$. Putting this all together, the equilibrium social loss takes the value in (5). We defer a proof of Proposition 2 to Appendix B.

3.3 Non-monotonicity along several axes of varying representations

Using Proposition 2, we next vary representations along several axes and compute the equilibrium social loss, observing non-monotonicity in each case.

Setting 1: Varying the per-representation Bayes risks. Consider a population with a single value of x that has Bayes risk $\alpha(x) = \alpha$. We vary representation quality by varying α from 0 to 0.5. Figure 2a depicts the result: by Proposition 2, the equilibrium social loss is zero if $\alpha > 1/m$ and is α if $\alpha < 1/m$, leading to non-monotonicity at $\alpha = 1/m$. When there are $m \geq 3$ model-providers, the equilibrium social loss is thus non-monotonic in α. (For $m = 2$, where $\alpha = 1/2$ is the maximum possible per-representation Bayes risk, the equilibrium social loss is monotone in α.) As the number of model-providers increases, the non-monotonicity occurs at a higher data representation quality (a lower Bayes risk).

[2] When $\alpha(x) = 1/m$, there turn out to be multiple pure-strategy equilibria with different social losses.

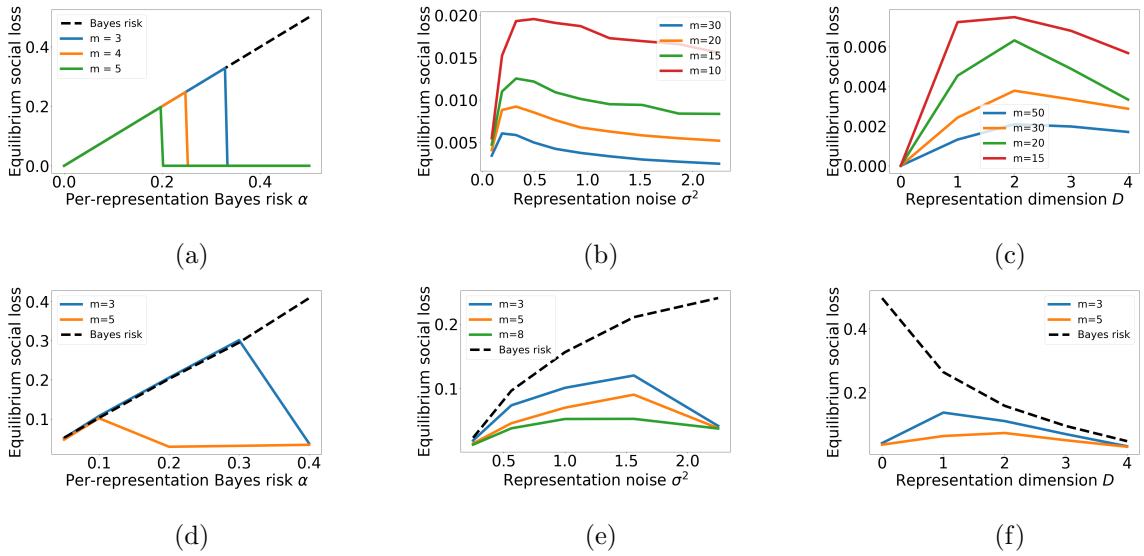


Figure 2: Equilibrium social loss (y-axis) versus data representation quality (x-axis) given m model-providers, for different function classes $\mathcal{F}$ (rows) and when representations are varied along different aspects (columns). Top row: $\mathcal{F} = \mathcal{F}_{\text{all}}^{\text{binary}}$, with closed-form formula from Proposition 2. Bottom row: linear functions, computed via simulation (Section 4). We vary representations with respect to per-representation Bayes risk (a,d), noise level (b,e), and dimension (c,f). The dashed line indicates the Bayes risk (omitted if it is too high to fit on the axis). The Bayes risk is monotone, but the equilibrium social loss is non-monotone.

Setting 2: Varying the representation noise. Consider a one-dimensional population given by a mixture of two Gaussians (one for each class), where each Gaussian has variance σ^2 (see Appendix A for the details of the setup). We vary the parameter σ to change the representation quality. Intuitively, a lower value of σ makes the Gaussians more well-separated, which improves representation quality (Bayes risk). By Proposition 2, the equilibrium social loss is $\mathbb{E}\left[\alpha(x) \cdot \mathbb{1}[\alpha(x) < 1/m]\right]$. For each value of σ, we estimate the equilibrium social loss by sampling representations x from the population and taking an average.[3] Figure 2b depicts the result: the equilibrium social loss is non-monotonic in σ (and thus the Bayes risk). Again, as the number of model-providers increases, the non-monotonicity occurs at a higher representation quality (a lower Bayes risk).

Setting 3: Varying the representation dimension. We consider a four-dimensional population (X^{all}, Y), and let the representation X consist of the first D coordinates of X^{all}, for D varying from 0 to 4 (see Appendix A for full details). Intuitively, a higher dimension D makes the representations more informative, thus improving representation quality (Bayes risk). As before, for each value of D, we estimate the equilibrium social loss by sampling representations x from the population and taking an average. Figure 2c depicts the result: the equilibrium social loss is once again non-monotonic in the representation dimension D (and thus the Bayes risk).

Discussion. Settings 1-3 illustrate that equilibrium social loss can be non-monotonic in Bayes risk when representations are improved along many qualitatively different axes. The intuition is

[3]Strictly speaking, we can't directly apply Proposition 2 to this setup since X is infinite. We circumvent this issue by applying Proposition 2 on a sample of the representations.

that varying representations along these axes can increase the values of $\alpha(x)$ for inputs x; by Proposition 2, these changes to $\alpha(x)$ can lead to non-monotonicity in the equilibrium social loss. We will revisit Settings 1-3 for richer market structures (Section 3.5) and for linear predictors and noisy user decisions (Section 4.2).

3.4 Generalization to more than 2 classes

While our analysis has thus far focused on classification with $K = 2$ classes, the number of classes K can be much larger in practice. As a motivating example, consider content recommendation tasks where each class represents a different genre of content; since the content landscape can be quite diverse, we would expect K to be fairly large.[4] This motivates us to extend our theoretical characterization in Proposition 2 to classification with $K \geq 2$ classes.

For the case of $K \geq 2$ classes, the appropriate analogue of the per-representation Bayes risk is the per-class-per-representation Bayes risk, defined to be:

$$\alpha^i(x) := \mathbb{P}(y = i \mid x) \tag{6}$$

for each $x \in X$ and $i \in \{0, 1, \ldots, K-1\}$. Observe that $1 - \max_{0 \leq i \leq K-1} \alpha^i(x)$ is the minimum error that a single model-provider can hope to achieve on x, and $\texttt{OPT}_{\text{single}}$ is equal to the average value $\mathbb{E}[1 - \max_{0 \leq i \leq K-1} \alpha^i(x)]$ across the population. The equilibrium social loss, however, depends on other aspects of the $\alpha^i(x)$ values.

We characterize the equilibrium social loss in terms of the per-class-per-representation Bayes risks in the following proposition. Our characterization again focuses on pure-strategy equilibria, which are guaranteed to exist in this setup by Proposition 1.

Proposition 3. *Let X be a finite set, let there be $K \geq 2$ classes, let $\mathcal{F} = \mathcal{F}_{all}^{multi\text{-}class}$. Suppose that user decisions are noiseless (i.e., user decisions are given by (3)). Let $c = \min_{x \in X} \max_{0 \leq i \leq K-1} \alpha^i(x)$. Then, at any pure strategy Nash equilibrium $f_1^*, \ldots, f_m^*$, the social loss $\texttt{SL}(f_1^*, \ldots, f_m^*)$ is bounded as*

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sum_{i=1}^{K} \alpha^i(x) \cdot \mathbb{1}\left[\alpha^i(x) < \frac{c}{m}\right] \right] \leq \texttt{SL}(f_1^*, \ldots, f_m^*) \leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sum_{i=1}^{K} \alpha^i(x) \cdot \mathbb{1}\left[\alpha^i(x) \leq \frac{1}{m}\right] \right]. \tag{7}$$

The high-level intuition for Proposition 3 is similar to the intuition for Proposition 2, except that each class needs to be considered separately. In particular, when class i occurs sufficiently frequently for the representation x (i.e., when $\alpha^i(x)$ is not too small), then some model-provider will label x as i; on the other hand, if the class i occurs very infrequently for x, then no model-provider will label x as i. We defer a proof of Proposition 3 to Appendix B.

While Proposition 3 is conceptually a generalization of Proposition 2, the details of Proposition 3 slightly differ. In particular, Proposition 3 does not completely pin down the equilibrium social loss, and there is a factor of c slack in the constraint on each $\alpha^i(x)$ in (7) between the upper and lower bounds. Nonetheless, since the value $c = \min_{x \in X} \max_{0 \leq i \leq K-1} \alpha^i(x)$ measures the minimum accuracy of the Bayes optimal predictor across all inputs x, we expect that "reasonable" representations (i.e., representations which are sufficiently informative) would have c equal to a constant. When c is a constant, there is at most a constant factor slack in the $\alpha^i(x)$ constraints in (7) between upper and lower bound.

[4] When K is large, even if users can "search" for and "consume" content on their own without relying on model-provider predictions, we expect that our measure of social loss would still be a good proxy for the loss experienced by users. In particular, it would be prohibitively expensive for users to try out all K classes, so classes that are not suggested to the user by any model-provider's predictions might be effectively inaccessible to the user.

For similar reasons to Proposition 2, Proposition 3 implies that the equilibrium social loss can be non-monotonic in the representation quality (i.e., the Bayes risk). As a concrete example, consider the following adaptation of Setting 1 in Section 3.2: let there be a population with a single value of x where $\alpha_0(x) = 1 - 2\alpha$, $\alpha_1(x) = \alpha$, and $\alpha_2(x) = \alpha$ for some $\alpha < 1/4$. In this setup, we see that $c \geq 1/2$. By Proposition 3, the equilibrium social loss is 2α if $\alpha < 1/(2m)$, and the equilibrium social loss is 0 if $\alpha > 1/m$; on the other hand, the Bayes risk is equal to 2α for any $\alpha < 1/4$. This illustrates that the equilibrium social loss is non-monotonic in the Bayes risk. We expect that other setups similar to those in Section 3.2 will also lead to non-monotonicity for multi-class tasks.

3.5 Generalization to unequal market reputations

While we assumed above that users evenly break ties between model-providers, in reality, users might be more likely to choose model-providers with a higher market reputation (e.g., established, popular model-providers). This motivates us to incorporate market reputations into user decisions.

Formally, we assign to each model-provider j a *market reputation* w_j, and we replace the logit model in (1) with a weighted logit variant. When $c \to 0$, rather than breaking ties uniformly, they are instead broken proportionally to w_j:

$$\mathbb{P}[j^*(x,y) = j] = \begin{cases} 0 & \text{if } j \notin \arg\min_{j' \in [m]} \mathbb{1}[y \neq f_{j'}(x)] \\ \frac{w_j}{\sum_{j'' \in [m]} w_{j''} \cdot \mathbb{1}[j'' \in \arg\min_{j' \in [m]} \mathbb{1}[y \neq f_{j'}(x)]]} & \text{if } j \in \arg\min_{j' \in [m]} \mathbb{1}[y \neq f_{j'}(x)]. \end{cases} \tag{8}$$

See Section 5 for further discussion of this model. For simplicity, we assume that market reputations are normalized to sum to one.

Similarly to Proposition 2, we derive a closed-form formula for the equilibrium social loss, focusing on the case of binary classification with $m = 2$ model-providers for analytic tractability. We observe non-monotonicity as before, but with a more complex functional form.

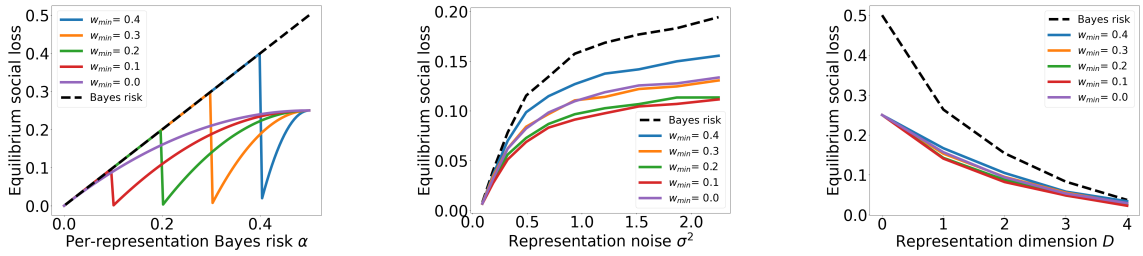
Proposition 4. *Let X be a finite set, let $K = 2$, and let $\mathcal{F} = \mathcal{F}_{all}^{binary}$. Suppose there are $m = 2$ model-providers with market reputations w_{min} and w_{max}, where $w_{max} \geq w_{min}$ and $w_{max} + w_{min} = 1$. Suppose that user decisions are given by (8), and that $\alpha(x) \neq w_{min}$ for all $x \in X$.[5] At any (mixed) Nash equilibrium (μ_1, μ_2), the expected social loss is equal to:*

$$\mathop{\mathbb{E}}_{\substack{f_1 \sim \mu_1 \\ f_2 \sim \mu_2}} [\mathit{SL}(f_1, f_2)] = \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} \left[\underbrace{\frac{(\alpha(x) - w_{min}) \cdot (w_{max} - \alpha(x))}{(1 - 2 \cdot w_{min})^2}}_{(A)} \cdot \mathbb{1}[\alpha(x) > w_{min}] + \underbrace{\alpha(x)}_{(B)} \cdot \mathbb{1}[\alpha(x) < w_{min}] \right]. \tag{9}$$

The high-level intuition for Proposition 4, like for Proposition 2, is that the equilibrium predictions go from heterogeneous to homogenous as $\alpha(x)$ decreases. Term (A), which is realized for large $\alpha(x)$, captures the equilibrium social loss for heterogeneous predictions. Term (B), which is realized for small $\alpha(x)$, captures the equilibrium social loss for homogeneous predictions. We defer the proof of Proposition 4 to Appendix B.

The details of Proposition 4 differ from Proposition 2 in several ways. First, the transition point from heterogeneous to homogeneous predictions occurs at $\alpha(x) = w_{\min}$ as opposed to $\alpha(x) = 1/2$. In particular, the transition point depends on the market reputations rather than only the

[5] As with Proposition 2, when $\alpha(x)$ is equal to $w_{\min}$ for some value of x, there are multiple equilibria.



(a) (b) (c)

Figure 3: Equilibrium social loss (y-axis) versus data representation quality (x-axis) given two model-providers with market reputations $[1 - w_{\min}, w_{\min}]$ when representations are varied along different aspects (columns). The equilibrium social loss is computed via the closed-form formula from Proposition 4. We vary representations with respect to per-representation Bayes risk (a), noise level (b), and dimension (c). The dashed line indicates the Bayes risk. The Bayes risk is monotone for all 3 axes of varying representations; on the other hand, the equilibrium social loss is non-monotone in the per-representation Bayes risk and monotone in noise level and dimension.

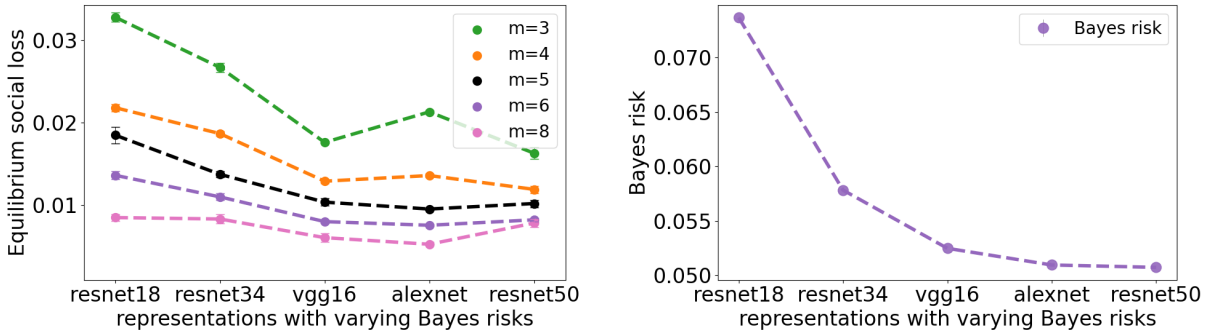
number of model-providers. Second, the equilibria have *mixed strategies* rather than pure strategies, because pure-strategy equilibria do not necessarily exist when market reputations are unequal (see Lemma 3 in Appendix B). Third, the social loss at a representation x is no longer equal to zero for heterogeneous predictions—in particular, term (A) is now positive for all $\alpha(x) > w_{\min}$ and increasing in $\alpha(x)$.

To better understand the implications of Proposition 4, we revisit Settings 1-3 from Section 3.3, considering the same three axes of varying representations with the same distributions over (x, y). In contrast to Section 3.3, we consider 2 competing model-providers with unequal market positions rather than m competing model providers with equal market positions. Our results, described below, are depicted in Figure 3.

Setting 1: Varying the per-representation Bayes risks. Consider the same setup as Setting 1 in Section 3.3. Figure 3a depicts the non-monotonicity of the equilibrium social loss in the per-representation Bayes risk α across different settings of market reputations for 2 competing model-providers. The discontinuity occurs at the smaller market reputation $w_{\min}$. Thus, as the market reputations of the 2 model-providers become closer together, the non-monotonicity occurs at a lower data representation quality (higher Bayes risk).

Settings 2-3: Varying the representation noise or representation dimension. Consider the setups from Settings 2-3 in Section 3.3. Figures 3b-3c depicts that the equilibrium social loss is *monotone* in data representation quality (Bayes risk) across different settings of market reputations for 2 competing model-providers.

Discussion. To interpret these results, observe that for 2 model-providers with equal market reputations ($w_{\min} = 0.5$), the equilibrium social loss is always equal to the Bayes risk by Propositions 2-4, which trivially implies monotonicity. In contrast, Figure 3 shows that for unequal market positions ($w_{\min} < 0.5$), the equilibrium social loss is non-monotonic in Bayes risk for Setting 1, though it is still monotonic in Bayes risk for Settings 2 and 3. (For comparison, recall from Figures 2a-2c that for $m \gg 2$ model-providers with equal market reputations, non-monotonicity



(a) (b)

Figure 4: Equilibrium social loss (left) and Bayes risk (right) on a binary classification task on CIFAR-10 (Section 4.3). Representations are generated from different networks pre-trained on ImageNet. The points show the equilibrium social loss when m model-providers compete with each other (left) and the Bayes risk of a single model-provider in isolation (right). While Bayes risk is decreasing in this representation ordering, the equilibrium social loss is non-decreasing in this ordering. The equilibrium social loss is thus non-monotonic in representation quality as measured by Bayes risk. Error bars are 1 standard error.

was exhibited for all three settings.) An interesting open question is identify other axes of varying representations, beyond Setting 1, which lead to non-monotonicity for 2 model-providers with unequal market reputations.

4 Empirical Analysis of Non-monotonicity for Linear Predictors

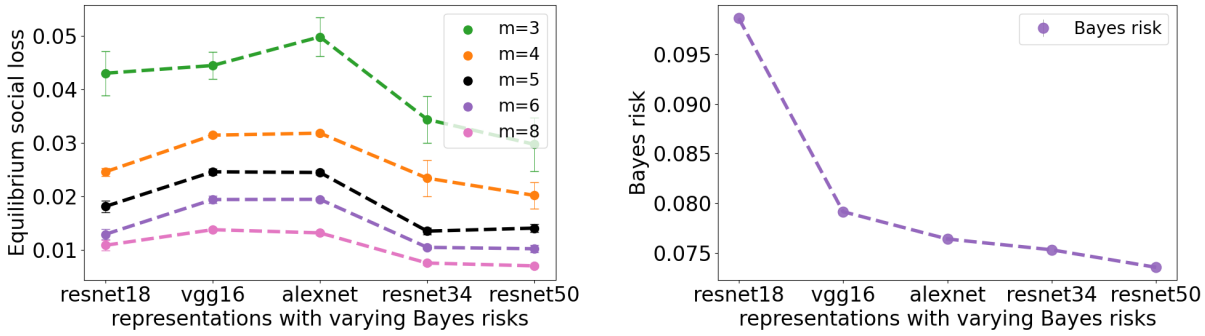
We next turn to linear predictors and demonstrate empirically that the social welfare can be non-monotonic in data representation quality in this setup as well.[6] We take $X = \mathbb{R}^D$ and we let the model parameters be ϕ. For binary classification, we let $\mathcal{F}_{\text{linear}}^{\text{binary}}$ be the family of linear predictors $f_{w,b} = \text{sigmoid}(\langle w, x\rangle + b)$ where $w \in \mathbb{R}^D$, $b \in \mathbb{R}$, and $\phi = [w, b]$. Similarly, for classification with more than 2 classes, we let $\mathcal{F}_{\text{linear}}^{\text{multi-class}}$ be the family of linear predictors $f_{W,b} = \text{softmax}(Wx + b)$ where $w \in \mathbb{R}^{|Y|\times D}$, $b \in \mathbb{R}^{|Y|}$, and $\phi = [W, b]$. Since this setting no longer admits closed-form formulae, we numerically estimate the equilibria using a variant of *best-response dynamics*, where model-providers repeatedly best-respond to the other predictors.

We first show on low-dimensional synthetic data on a binary classification task that the insights from Section 3.3 readily generalize to linear predictors (see Figures 2d-2f). We then turn to natural data, considering binary and 10-class image classification tasks for CIFAR-10 and using pretrained networks—AlexNet, VGG16, and various ResNets—to generate high-dimensional representations (ranging from 512 to 4096). In this setting we again find that the equilibrium social loss can be non-monotonic in the Bayes risk (see Figure 4 and Figure 5).

4.1 Best-response dynamics implementation

To enable efficient computation, we assume the distribution $\mathcal{D}$ corresponds to a finite dataset with N data points. We calculate equilibria using an approximation of best-response dynamics. Model-

[6] The code can be found at https://github.com/mjagadeesan/competition-nonmonotonicity.



(a) (b)

Figure 5: Equilibrium social loss (left) and Bayes risk (right) on a 10-class classification task on CIFAR-10 (Section 4.4). Representations are generated from different networks pre-trained on ImageNet. The points show the equilibrium social loss when m model-providers compete with each other (left) and the Bayes risk of a single model-provider in isolation (right). While Bayes risk is decreasing in this representation ordering, the equilibrium social loss is non-decreasing in this ordering. The equilibrium social loss is thus non-monotonic in representation quality as measured by Bayes risk. Error bars are 1 standard error.

providers (players) iteratively (and approximately) best-respond to the other players' actions. We implement the approximate best-response as running several steps of gradient descent.

In more detail, for each $j \in [m]$, we initialize the model parameters ϕ as mean zero Gaussians with standard deviation σ. Our algorithm then proceeds in stages. At a given stage, we iterate through the model-providers in the order $1, \ldots, m$. When j is chosen, first we decide whether to reinitialize: if the risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_\phi(x), y)]$ exceeds a threshold ρ, we re-initialize w_j and b_j (sampling from mean zero Gaussians as before); otherwise, we do not reinitialize. Then we run gradient descent on $u(\cdot; \mathbf{f}_{-j})$ (computing the gradient on the full dataset of N points) with learning rate η for I iterations, updating the parameters ϕ. We run this gradient descent step up to 2 more times if the risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_\phi(x), y)]$ exceeds a threshold ρ'. At the end of a stage, the stopping condition is that for every $j \in [m]$, model-provider j's utility $u(f_j, \mathbf{f}_{-j})$ has changed by at most ϵ relative to the previous stage. If the stopping condition is not met, we proceed to the next stage.

4.2 Simulations on synthetic data

We first revisit Settings 1-3 from Section 3.3, considering the same three axes of varying representations with the same distributions over (x, y). In contrast to Section 3.3, we restrict the model family to linear predictors $\mathcal{F}^{\text{binary}}_{\text{linear}}$ instead of allowing all predictors $\mathcal{F}^{\text{binary}}_{\text{all}}$. We also set the noise parameter c in user decisions (1) to 0.3. Our goal is to examine if the findings from Section 3 generalize to this new setting.

We compute the equilibria for each of the following (continuous) distributions as follows. First, we let $\mathcal{D}$ be the empirical distribution over $N = 10,000$ samples from the continuous distribution. Then we run the best-response dynamics described in Section 4.1 with $\rho = 0.3$, $I = 5000$, $\eta = 0.1$, $\epsilon = 0.01$, and $\sigma = 0.1$. We then compute the equilibrium social loss according to (2). We also compute the Bayes optimal predictor with gradient descent. See Appendix A for full details.

Our results, described below, are depicted in Figures 2d-2f (row 2). We compare these results with Figures 2a-2c (row 1), which shows the analogous results for $\mathcal{F}^{\text{binary}}_{\text{all}}$ from Section 3.3.

Setting 1: Varying the per-representation Bayes risks. Consider the same single x setup as in Setting 1 in Section 3.3. The only parameter of the predictor is the bias $b \in \mathbb{R}$ (i.e., we treat x as zero-dimensional). Figure 2d shows that the equilibrium social loss is non-monotonic in α, which mirrors the non-monotonicity in Figure 2a.

Setting 2: Varying the representation noise. Consider the same one-dimensional mixture-of-Gaussians distribution as in Setting 2 in Section 3.3. (The weight w is one-dimensional.) We again vary the noise σ to change the representation quality. Figure 2e shows that the equilibrium social loss is non-monotonic in the noise σ, which again mirrors the non-monotonicity in Figure 2b.

Setting 3: Varying the representation dimension. Consider the same four-dimensional population as in Setting 3 in Section 3.3. We vary the representation dimension D from 0 to 4 to change the representation quality. Figure 2f shows that the equilibrium social loss is non-monotonic in the dimension D, which once again mirrors the non-monotonicity in Figure 2c.

Discussion. In summary, in Figure 2, rows 1 and 2 exhibit similar non-monotonicities. This illustrates that the insights from Section 3.2 translate to linear predictors and noisy user decisions.

4.3 Simulations on CIFAR-10 for binary classification

We next turn to experiments with natural data. While we have directly varied the informativeness of data representations thus far, representations in practice are frequently generated by pretrained models. The choice of the pretrained model implicitly influences representation quality, as measured by Bayes risk on the downstream task. In this section, we consider how the equilibrium social loss changes with representations generated from pretrained models of varying quality. We restrict the model family to linear predictors $\mathcal{F}_{\text{linear}}^{\text{binary}}$ and set the noise parameter c in user decisions (1) to 0.1.

We consider a binary image classification task on CIFAR-10 [Kri09] with 50,000 images. Class 0 is defined to be $\{\text{airplane}, \text{bird}, \text{automobile}, \text{ship}, \text{horse}, \text{truck}\}$ and the class 1 is defined to be $\{\text{cat}, \text{deer}, \text{dog}, \text{frog}\}$. We treat the set of 50,000 images and labels as the population of users, meaning that it is both the training set and the validation set.[7] Representations are generated from five models—AlexNet [KSH12], VGG16 [SZ15], ResNet18, ResNet34, and ResNet50 [HZR+16]—pretrained on ImageNet [DDS+09]. The representation dimension is 4096 for AlexNet and VGG16, 512 for ResNet18 and ResNet34, and 2048 for ResNet50.

We compute the equilibria as follows. First, we let $\mathcal{D}$ be the distribution described above with $N = 50,000$ data points. Then we run the best-response dynamics described in Section 4.1 for $m \in \{3, 4, 5, 6, 8\}$ model-providers with $\rho = \rho' = 0.3$, $I = 2000$, $\epsilon = 0.001$, $\sigma = 0.5$, and a learning rate schedule that starts at $\eta = 1.0$. We then compute the equilibrium social loss according to (2). We also compute the Bayes risk using gradient descent. For full experimental details, see Appendix A.

Figure 4 shows that the equilibrium social loss can be non-monotone in the Bayes risk. For example, for $m = 3$, VGG16 outperforms AlexNet, even though the Bayes risk of VGG16 is substantially higher than the Bayes risk of AlexNet. Interestingly, the location of the non-monotonicity differs across different values of m. For example, for $m = 5$ and $m = 8$, AlexNet outperforms ResNet50 despite having a higher Bayes risk, but ResNet50 outperforms AlexNet for $m = 3$ and $m = 4$.

[7] We make this choice to be consistent with the rest of the paper, where we focus on population-level behavior and thus do not consider generalization error.

4.4 Simulations on CIFAR-10 for 10-class classification

While our empirical analysis has thus far focused on binary classification, we now turn to classification with more than 2 classes. In particular, we consider a ten class CIFAR-10 [Kri09] task with 50,000 images. The labels are specified by the CIFAR-10 classes in the original dataset. We treat the set of 50,000 images and labels as the population of users, meaning that it is both the training set and the validation set. Representations are generated from the same five models—AlexNet [KSH12], VGG16 [SZ15], ResNet18, ResNet34, and ResNet50 [HZR+16]—pretrained on ImageNet [DDS+09]. We restrict the model family to linear predictors $\mathcal{F}_{\text{linear}}^{\text{multi-class}}$ and again set the noise parameter c in user decisions (1) to 0.1.

We compute the equilibria as follows. First, we let $\mathcal{D}$ be the distribution described above with $N = 50,000$ data points. Then we run the best-response dynamics described in Section 4.1 for $m \in \{3, 4, 5, 6, 8\}$ model-providers with $\rho = 0.7$, $\rho' = 1.0$, $I = 2000$, $\epsilon = 0.001$, $\sigma = 0.5$, and a learning rate schedule that starts at $\eta = 1.0$. As before, we compute the equilibrium social loss according to (2), and we also compute the Bayes risk using gradient descent. For full experimental details, see Appendix A.

Figure 5 shows that the equilibrium social loss can be non-monotone in the Bayes risk. For example, across all five values of m, ResNet18 outperforms VGG16, even though the Bayes risk of ResNet is substantially higher than the Bayes risk of VGG16. Furthermore, for $m = 3$, VGG16 outperforms AlexNet despite having a larger Bayes risk. Interestingly, the shape of the equilibrium social loss curve for each value of m (Figure 5a) appears qualitatively different than the analogous equilibrium social loss curve for binary classification (Figure 4a).

5 Discussion of Model Assumptions

We highlight and discuss several assumptions that we make in our stylized model.

5.1 Assumptions on user decisions

Our primary model for user decisions given by (1) is the standard logit model for discrete choice decisions [Tra02] which is also known as the Boltzmann rationality model. In the limit as $c \to 0$, a user with representation x and label y select from the set of model-providers $\arg\min_{j \in [m]} \ell(f_j(x), y)$ that achieve the minimum loss; in particular, the user chooses a model-provider from this set with probability proportional to the model-provider's market reputation. For $c > 0$, the specification in equation (1) captures that users evaluate a model-provider based on a noisy perception of the loss.

While this model implicitly assumes that a user's choice of platform is fully specified by the platforms' choices of predictor (i.e. platforms are ex-ante homogeneous), we extend this model in Section 3.5 to account for uneven market reputations across decisions. These market reputations are modeled as global weights in the logit model for discrete choice. Given market reputations $w_1, \ldots, w_m$, users choose a predictor according to:

$$\mathbb{P}[j^*(x, y) = j] = \frac{w_j \cdot e^{-\ell(f_j(x), y)/c}}{\sum_{j'=1}^{m} w_{j'} \cdot e^{-\ell(f_{j'}(x), y)/c}}. \tag{10}$$

When the market reputations are all equal ($w_1 = \ldots = w_m$), equation (10) exactly corresponds to (1). When the market reputations w_j are not equal, equation (10) captures that users place a higher weight on model-providers with a higher market reputation. This captures that users are more likely to choose a popular model-provider than a very small model-provider without much

reputation. However, this formalization does assume that market reputations are global across users and that market reputations surface as tie-breaking weight in the noiseless limit.

Implicit in this model is asymmetric information between the model-providers and users. While the only information that a model-provider has about users is their representations, a user can make decisions based on noisy perceptions of their own loss (which can depend on their label). This captures that, even if users are unlikely to know their own labels, users can experiment with multiple model-providers to (noisily) determine which one maximizes their utility. The inclusion of market reputations reflects that users are more likely to experiment with and ultimately choose popular model-providers than less popular model-providers.

5.2 Assumption of global data representations

Our results assume that all model-providers share the same representations x for each user and thus improvements in representations x are experienced by all model-providers. This assumption is motivated by emerging marketplaces where different model-providers utilize the same pretrained model, but *finetune* the model in different ways. To simplify this complex training process, we conceptualize pretraining as *learning data representations (e.g., features)* and fine-tuning as *learning a predictor from these representations*. In this formalization, increasing the scale of the pretrained model (e.g., by increasing the number of parameters or the amount of data) leads to improvements in data representations accessible to all of the model-providers during "fine-tuning".

An interesting direction for future work would be to incorporate heterogeneity or local improvements in the data representations.

5.3 Assumption on model-provider action space

We make the simplifying assumption that the only action taken by model-providers is to choose a classifier from a pre-specified class. This formalization does not capture other actions (such as data collection and price setting) that may be taken by the platform. Incorporating other model-provider decisions would be an interesting avenue for future work.

6 Discussion

We showed that the monotonicity of scaling trends can be violated under competition. In particular, we demonstrated that when multiple model-providers compete for users, improving data representation quality (as measured by Bayes risk) can *increase* the overall loss at equilibrium. We exhibited the non-monotonicity of the equilibrium social loss in the Bayes risk when representations are varied along several axes (per-representation Bayes risk, noise, dimension, and pre-trained model used to generate the representations).

An interesting direction for future work is to further characterize the regimes when the equilibrium social loss is monotonic versus non-monotonic in data representation quality as measured by Bayes risk. For example, an interesting open question is to generalize our theoretical results from Section 3 to more general function classes and distributions of market reputations. Moreover, another interesting direction would be generalize our empirical findings from Section 4 to other axes of varying data representations and to non-linear classes of predictors. Finally, while we have focused on classification tasks, it would be interesting to generalize our findings to regression tasks with continuous outputs or to generative AI tasks with text-based or image-based outputs.

More broadly, the non-monotonicity of equilibrium social welfare in scale under competition establishes a disconnect between scaling trends in the single model-provider setting and in the com-

petitive setting. In particular, typical scaling trends (e.g. [KMH+20; SK20; BDK+21; HBM+22; HKH+21])—which show increasing scale reliably increases predictive accuracy for a single model-provider in isolation—may not translate to competitive settings such as digital marketplaces. Thus, understanding the downstream impact of scale on user welfare in digital marketplaces will likely require understanding how scaling trends behave under competition. We hope that our work serves as a starting point for analyzing and eventually characterizing the scaling trends of learning systems in competitive settings.

7 Acknowledgments

We thank Yiding Feng, Xinyan Hu, and Alex Wei for useful comments on the paper. This work was partially supported by the Open Phil AI Fellowship, the Berkeley Fellowship, the European Research Council (ERC) Synergy Grant program, and the National Science Foundation under grant numbers 2031899 and 1804794.

References

[AdPT92] Simon P. Anderson, Andre de Palma, and Jacques-Francois Thisse. *Discrete Choice Theory of Product Differentiation*. The MIT Press, October 1992.

[AMS+20] Guy Aridor, Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: the perils of exploration under competition. *CoRR*, abs/2007.10144, 2020.

[BCK+22] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: does algorithmic monoculture lead to outcome homogenization? In *NeurIPS*, 2022.

[BCV13] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[BDK+21] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *CoRR*, abs/2102.06701, 2021.

[BKS12] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *JMLR*, 13(1):2617–2654, 2012.

[BT17] Omer Ben-Porat and Moshe Tennenholtz. Best response regression. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1499–1508, 2017.

[BT19] Omer Ben-Porat and Moshe Tennenholtz. Regression equilibrium. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, pages 173–191. ACM, 2019.

[CDR+22] Micah D. Carroll, Anca D. Dragan, Stuart Russell, and Dylan Hadfield-Menell. Estimating and penalizing induced preference shifts in recommender systems. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2686–2708. PMLR, 2022.

[CHR+22] Mihaela Curmei, Andreas A. Haupt, Benjamin Recht, and Dylan Hadfield-Menell. Towards psychologically-grounded dynamic preference models. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, pages 35–48, 2022.

[CP21] Emilio Calvano and Michele Polo. Market power, competition and innovation in digital markets: a survey. *Information Economics and Policy*, 54:100853, 2021.

[CRV+04] Shih-fen Cheng, Daniel M. Reeves, Yevgeniy Vorobeychik, and Michael P. Wellman. Notes on equilibria in symmetric games. In *Proceedings of the 6th International Workshop on Game Theoretic and Decision Theoretic Agents (GTDT)*, pages 71–78, 2004.

[DCR+22] Sarah Dean, Mihaela Curmei, Lillian J. Ratliff, Jamie Morgenstern, and Maryam Fazel. Multi-learner risk reduction under endogenous participation dynamics. *CoRR*, abs/2206.02667, 2022.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[dGT79] C. d'Aspremont, J. Jaskold Gabszewicz, and J.-F. Thisse. On hotelling's "stability in competition". *Econometrica*, 47(5):1145–1150, 1979. ISSN: 00129682, 14680262.

[DM22] Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations. In David M. Pennock, Ilya Segal, and Sven Seuken, editors, *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, pages 795–816. ACM, 2022.

[FGH+19] Yiding Feng, Ronen Gradwohl, Jason D. Hartline, Aleck C. Johnsen, and Denis Nekipelov. Bias-variance games. *CoRR*, abs/1909.03618, 2019.

[GZK+21] Tony Ginart, Eva Zhang, Yongchan Kwon, and James Zou. Competing AI: how does competition feedback affect machine learning? In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 1693–1701, 2021.

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022.

[HKH+21] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

[HMP+16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 7th Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 111–122, 2016.

[Hot81] Harold Hotelling. Stability in competition. *Economic Journal*, 39(153):41–57, 1981.

[HSN+18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, October 2018.

[HZR+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[IK22] Ganesh Iyer and T. Tony Ke. Competitive algorithmic targeting and model selection. Available at SSRN: https://ssrn.com/abstract=4214973, 2022.

[IKL+11] Nicole Immorlica, Adam Tauman Kalai, Brendan Lucier, Ankur Moitra, Andrew Postlewaite, and Moshe Tennenholtz. Dueling algorithms. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 215–224, 2011.

[JJH22] Meena Jagadeesan, Michael I. Jordan, and Nika Haghtalab. Competition, alignment, and equilibria in digital marketplaces. *CoRR*, abs/2208.14423, 2022.

[JS21] Bruno Jullien and Wilfried Sand-Zantman. The economics of platforms: a theory guide for competition policy. *Information Economics and Policy*, 54:100880, 2021.

[KGZ22] Yongchan Kwon, Tony Ginart, and James Zou. Competition over data: how does data purchase affect users? *Trans. Mach. Learn. Res.*, 2022.

[KMH+20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

[KR21] Jon M. Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proc. Natl. Acad. Sci. USA*, 118(22):e2018340118, 2021.

[Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[LDR+18] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3156–3164. PMLR, 2018.

[LWH+20] Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer T. Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 381–391. ACM, 2020.

[MS96] Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996.

[Ros73] Robert W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *Int. J. Game Theory*, 2(1):65–67, December 1973.

[SK20] Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *CoRR*, abs/2004.10802, 2020.

[SZ15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Tra02] Kenneth Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, New York, 2002.

A Additional Details of Simulations

Hyperparameters. We introduce a temperature parameter τ within our loss function, defining the loss $\ell(f_{w,b}(x), y)$ to be $|\operatorname{sigmoid}((\langle w, x\rangle + b)/\tau) - 1|$. This reparameterizes, but does not change, the model family.

When we run the best-response dynamics, we always initialize the model parameters as mean-zero Gaussians with standard deviation σ. When we reinitialize model parameters, we again initialize them as mean-zero Gaussians with standard deviation σ. For Section 4.2, we set $I = 5000$, $\tau = 0.1$, $\epsilon = 0.001$, $\eta = 0.1$, $\sigma = 0.1$, and $\rho = \rho' = 1.0$. For Section 4.3 and Section 4.4, we set $I = 2000$, $\sigma = 0.5$, $\tau = 1.0$, $\epsilon = 0.001$, and η with the following learning rate schedule to expedite convergence: $\eta = 1.0$ if the risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_{w_j,b_j}(x), y)]$ is at least 0.5, $\eta = 5.0$ if the risk is in $[0.4, 0.5)$, $\eta = 15$ if the risk is in $[0.3, 0.4)$, and $\eta = 20$ if the risk is less than 0.3. We set $\rho = \rho' = 0.3$ for Section 4.3 and we set $\rho = 0.7$ and $\rho' = 1$ for Section 4.4.

For Section 4.3 and Section 4.4, we ran over several trials for each data point and the error bars show two standard errors from the mean. For binary classification, the number of trials was 20 for $m = 3$ and $m = 4$ and 8 for $m = 5$, $m = 6$, and $m = 8$. For 10-class classification, the number of trials was 40 for $m = 3$ and $m = 4$ and 8 for $m = 5$, $m = 6$, and $m = 8$.

In addition to computing the equilibria, we also approximate the optimal Bayes risk. For Section 4.2, we run gradient descent for $10,000$ iterations with learning rate equal to one and parameters initialized to independent Gaussians with zero mean and standard deviation 0.1. For Section 4.3, we run gradient descent for $50,000$ iterations with learning rate equal to 0.1 and parameters initialized to independent Gaussians with zero mean and standard deviation 0.005. For Section 4.4, we run gradient descent for $70,000$ iterations with learning rate equal to 0.1 and parameters initialized to independent Gaussians with zero mean and standard deviation 0.005.

Generation of the synthetic dataset. In Setting 1 (Figures 2a, 3a, and 2d), we consider a zero-dimensional population where $Y \mid X$ is distributed as a Bernoulli with probability α. In Figure 2d, the meaning of a zero-dimensional representation is that the only parameter is the bias.

In Setting 2 (Figures 2b, 3b, and 2e), we consider a one-dimensional population given by a mixture of Gaussians. In particular, the Gaussian $X \mid Y = 0$ is distributed as $N(-\mu, \sigma^2)$ and the Gaussian $X \mid Y = 1$ is distributed as $N(\mu, \sigma^2)$. The mean μ is taken to be 1. The distribution of the labels is given by $\mathbb{P}[Y = 1] = 0.4$ and $\mathbb{P}[Y = 1] = 0.6$.

In Setting 3 (Figures 2c, 3c, and 2f), let $D_{\text{base}} = 4$. The distribution over (X^{all}, Y) consists of D_{base} subpopulations. We define the distribution of (X^{all}, Y) as follows: each subpopulation $1 \le i \le D_{\text{base}}$ has a different mean vector $\mu_i \in \mathbb{R}^{D_{\text{base}}}$ and is distributed as $X^{\text{all}} \sim Y = 0 \sim N(-\mu_i, \sigma^2)$, let $X^{\text{all}} \sim Y = 1 \sim N(\mu_i, \sigma^2)$, and let $\mathbb{P}[Y = 0] = \mathbb{P}[Y = 1] = 1/2$.We define $(\mu_i)_d = 0$ for $1 \le d \le i - 1$ and $(\mu_i)_d = 1$ for $i \le d \le D_{\text{base}}$, and we let $\sigma = 1$. If the representation dimension is D, then we define X to consist of the first D coordinates of X^{all}. When $D = 0$, the model-provider is not given representations and thus must assign all users to the same output. (Our setup captures that the dimension D must be at least i to see any nontrivial features about subpopulation i.) The distribution across the 4 subpopulations is 0.7, 0.15, 0.1, and 0.05.

In each case, we draw 10,000 samples and take the resulting empirical distribution to be $\mathcal{D}$.

Generation of the CIFAR-10 task. We consider a binary classification task consisting of the first 10,000 images in the training set of CIFAR-10. The class 0 is defined to be {airplane, bird, automobile, ship, horse, truck} and class 1 is defined to be {cat, deer, dog, frog}. To generate representations, we use the pretrained models from the Pytorch `torchvision.models` package; these models were pretrained on ImageNet.

Compute details. We run our simulations on a single A100 GPU.

B Additional Results and Proofs for Section 3

In Appendix B.1, we show a decomposition lemma and prove existence of equilibrium (Proposition 1). We prove the results from Section 3.2 in Appendix B.2, prove the results from Section 3.4 in Appendix B.3, and prove the results from Section 3.5 in Appendix B.4.

B.1 Decomposition lemma and existence of equilibrium

We first show that we can decompose model-provider actions into independent decisions about each representation x. To formalize this, let $\mathcal{D}$ be the data distribution, and let $\mathcal{D}_x$ be the conditional distribution over $(X,Y) \mid X = x$ where $(X,Y) \sim \mathcal{D}$. Let $(\mathcal{F}_{\text{all}}^{\text{multi-class}})^x := \left\{f^0, f^1, \ldots, f^{K-1}\right\}$ be the class of K functions from a single representation x to $\{0,1,\ldots,K-1\}$, where $f^i(x) = i$.

Lemma 1. *Let X be a finite set of representations, let $\mathcal{F} = \mathcal{F}_{all}^{multi\text{-}class}$, and let $\mathcal{D}$ be the distribution over (X,Y). For each $x \in X$, let $\mathcal{D}_x$ be the conditional distribution over $(X,Y) \mid X = x$ where $(X,Y) \sim \mathcal{D}$, and let $(\mathcal{F}_{all}^{multi\text{-}class})^x := \left\{f^0, f^1, \ldots, f^{K-1}\right\}$ be the class of the K functions from a single representation x to $\{0,1\}$, where $f^i(x) = i$. Suppose that user decisions are noiseless (i.e., $c \to 0$, so user decisions are given by (3)). A market outcome $f_1, \ldots, f_m$ is a pure-strategy equilibrium if and only if for every $x \in X$, the market outcome $(f^{f_1(x)}, \ldots, f^{f_m(x)})$ is a pure-strategy equilibrium for $(\mathcal{F}_{all}^{multi\text{-}class})^x$ with data distribution $\mathcal{D}_x$.*

The intuition is that since $\mathcal{F}_{\text{all}}^{\text{multi-class}}$ is all possible functions, model-providers make independent decisions for each data representation.

Proof. Let $\mathcal{D}^R$ be the marginal distribution of X with respect to the distribution $(X,Y) \sim \mathcal{D}$. First, we write model-provider j's utility as:

$$u(f_j; \mathbf{f}_{-j}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathbb{P}[j^*(x,y) = j]\right] = \mathop{\mathbb{E}}_{x'\sim\mathcal{D}^R} \left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}} \left[\mathbb{P}[j^*(x,y) = j]\right]\right], \tag{11}$$

where $\mathbf{f}_{-j}$ denotes the predictors chosen by the other model-providers. The key intuition for the proof will be that the predictions $[f_1(x''), \ldots, f_m(x'')]$ affect $\mathbb{E}_{(x,y)\sim\mathcal{D}_{x'}} \left[\mathbb{P}[j^*(x,y) = j]\right]$ if and only if $x' = x''$.

First we show that if $f_1, \ldots, f_m$ is a pure-strategy equilibrium, then $(f^{f_1(x')}, \ldots, f^{f_m(x')})$ is a pure-strategy equilibrium for $(\mathcal{F}_{\text{all}}^{\text{multi-class}})^{x'}$ with data distribution $\mathcal{D}_{x'}$. Assume for sake of contradiction that $(f^{f_1(x')}, \ldots, f^{f_m(x')})$ is not an equilibrium. Then there exists $j' \in [m]$ such that model-provider j' would achieve higher utility if they switched from $f^{f_{j'}(x')}$ to f^l for some $l \neq f_{j'}(x')$. Let $f'_{j'}$ be the predictor given by $f'_{j'}(x) = f_{j'}(x)$ if $x \neq x'$ and $f'_{j'}(x') = l$. By equation (11), this would mean that $u(f'_{j'}; \mathbf{f}_{-j'})$ is strictly higher than $u(f_{j'}; \mathbf{f}_{-j'})$ which is a contradiction.

Next, we show that if $(f^{f_1(x')}, \ldots, f^{f_m(x')})$ is a pure-strategy equilibrium for $(\mathcal{F}_{\text{all}}^{\text{binary}})^{x'}$ with data distribution $\mathcal{D}_{x'}$ for all $x' \in X$ then $f_1, \ldots, f_m$ is a pure-strategy equilibrium. Assume for sake of contradiction that there exists j' such that $u(f'_{j'}; \mathbf{f}_{-j'}) > u(f_j; \mathbf{f}_{-j'})$. By equation (11), there must exist x' such that $\mathbb{E}_{(x,y)\sim\mathcal{D}_{x'}} \left[\mathbb{P}[j^*(x,y) = j']\right]$ is higher for $f'_{j'}$ than $f_{j'}$. This means that $(f^{f_1(x')}, \ldots, f^{f_m(x')})$ is not an equilibrium (since f^l would be a better response for model-provider j') which is a contradiction. □

We next prove Proposition 1, showing that a pure-strategy equilibrium exists by applying the proof technique of Lemma 3.7 of Ben-Porat and Tennenholtz [BT19].

Proof of Proposition 1. By Lemma 1, it suffices to show that there exists a pure-strategy equilibrium whenever there is a single data representation $X = \{x\}$. In this case, the function class $\mathcal{F}_{\text{all}}^{\text{multi-class}}$ consists of K predictors $\left\{f^0, f^1 \ldots, f^{K-1}\right\}$ given by $f^i(x) = i$. For each class i, let $\mathbb{P}[Y = i \mid X] = p_i$.

For the special case of $K = 2$ (binary classification), the game between model-providers is thus a 2-action game with symmetric utility functions. Thus, it must possess a (possibly asymmetric) pure Nash equilibrium [CRV+04].

For the general case of $K \geq 2$, we can no longer apply the result in [CRV+04] since there can be more than 2 actions. We instead show that the game is a potential game, following a similar argument to Ben-Porat and Tennenholtz [BT19]. We define the potential function $\Phi(\cdot)$ as follows. For each $i \in \{0, 1, \ldots, K-1\}$, we define the function $G_i : \left\{f^0, f^1 \ldots, f^{K-1}\right\}^m \to \mathbb{R}_{\geq 0}$ to be:

$$G_i(f_1, \ldots, f_m) := \begin{cases} \frac{1}{m} & \text{if } \left|\left\{j \in [m] \mid f_j = f^i\right\}\right| = 0 \\ \sum_{l=1}^{\left|\left\{j \in [m] \mid f_j = f^i\right\}\right|} \frac{1}{l} & \text{if } \left|\left\{j \in [m] \mid f_j = f^i\right\}\right| \geq 1. \end{cases}$$

We let

$$\Phi(f_1, \ldots, f_m) := \sum_{i=1}^{K} p_i \cdot G_i(f_1, \ldots, f_m).$$

We show that Φ is a potential function for this game. Suppose that model-provider j switches from $f_j := f^i$ to $f'_j = f^{i'}$ for $i' \neq i$. For each $i \in \{0, 1, \ldots, K-1\}$, let $N_i = \left|\left\{j \in [m] \mid f_j = f^i\right\}\right|$ be the number of model-providers who choose f^i on the original outcome $[f_1, \ldots, f_m]$. We observe that:

$$u(f_j; \mathbf{f}_{-j}) - u(f'_j; \mathbf{f}_{-j}) = \begin{cases} p_i \cdot \frac{1}{N_i} - p_{i'} \cdot \frac{1}{N_{i'}+1} & \text{if } N_i > 1, N_{i'} > 0 \\ p_i \cdot \left(1 - \frac{1}{m}\right) - p_{i'} \cdot \frac{1}{N_{i'}+1} & \text{if } N_i = 1, N_{i'} > 0 \\ p_i \cdot \frac{1}{N_i} - p_{i'} \cdot \left(1 - \frac{1}{m}\right) & \text{if } N_i > 1, N_{i'} = 0 \\ p_i \cdot \left(1 - \frac{1}{m}\right) - p_{i'} \cdot \left(1 - \frac{1}{m}\right) & \text{if } N_i = 1, N_{i'} = 0 \end{cases}$$

Moreover, we see that:

$$\begin{aligned} &\Phi(f_1, \ldots, f_m) - \Phi(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m) \\ &= \sum_{i''=1}^{K} p_{i''} \cdot G_{i''}(f_1, \ldots, f_m) - \sum_{i''=1}^{K} p_{i''} \cdot G_{i''}(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m) \\ &= p_i \cdot \left(G_i(f_1, \ldots, f_m) - G_i(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m)\right) \\ &+ p_{i'} \left(G_{i'}(f_1, \ldots, f_m) - G_{i'}(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m)\right). \end{aligned}$$

If $N_i > 1$, then:

$$G_i(f_1, \ldots, f_m) - G_i(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m) = \frac{1}{N^i}$$

and if $N_i = 1$, then

$$G_i(f_1, \ldots, f_m) - G_i(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m) = 1 - \frac{1}{m}.$$

Similarly, if $N_{i'} > 0$, then:

$$G_{i'}(f_1, \ldots, f_m) - G_{i'}(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m) = -\frac{1}{N^{i'} + 1}$$

and if $N_{i'} = 0$, then

$$G_{i'}(f_1, \ldots, f_m) - G_i(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m) = -\left(1 - \frac{1}{m}\right).$$

Altogether, this implies that:

$$\Phi(f_1, \ldots, f_m) - \Phi(f_1, f_2, \ldots, f_{j-1}, f'_j, f_{j+1}, \ldots, f_m) = u(f_j; \mathbf{f}_{-j}) - u(f'_j; \mathbf{f}_{-j}),$$

which shows that Φ is a potential function of the game. Since pure strategy equilibria exist in potential games [Ros73; MS96], a pure strategy equilibrium must exist in the game. □

B.2 Proofs for Section 3.2

We next prove Proposition 2. The high-level intuition of the proof is as follows. By Lemma 1, we can focus on one data representation x at a time. Let $y^* = \arg\max_y \mathbb{P}[y \mid x]$ be the Bayes optimal label of x. The proof boils down to characterizing when the market outcome, $f_j(x) = y^*$ for $j \in [m]$, is an equilibrium, and the equilibrium social loss is determined by whether this market outcome is an equilibrium or not.

Proof of Proposition 2. Let $\mathcal{D}^R$ be the marginal distribution of X with respect to the distribution $(X, Y) \sim \mathcal{D}$. Let $f_1^*, \ldots, f_m^*$ be a pure-strategy equilibrium. The social loss is equal to:

$$\begin{aligned}
\texttt{SL}(f_1^*, \ldots, f_m^*) &= \mathbb{E}[\ell(f^*_{j^*(x,y)}(x), y)] \\
&= \mathop{\mathbb{E}}_{x' \sim \mathcal{D}^R}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(f^*_{j^*(x,y)}(x), y) \mid x = x']\right] \\
&= \mathop{\mathbb{E}}_{x' \sim \mathcal{D}^R}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f^*_{j^*(x,y)}(x), y)]\right],
\end{aligned}$$

where $\mathcal{D}_{x'}$ denotes the conditional distribution $(X, Y) \mid X = x'$ where $(X, Y) \sim \mathcal{D}$. Thus, to analyze the overall social loss, we can separately analyze the social loss on each distribution $\mathcal{D}_{x'}$ and then average across distributions. It suffices to show that $\mathbb{E}_{\mathcal{D}_{x'}}[\ell(f^*_{j^*(x,y)}(x), y)] = \alpha(x')$ if $\alpha(x') < 1/m$ and zero if $\alpha(x') > 1/m$.

To compute the social loss on $\mathcal{D}_{x'}$, we first apply Lemma 1. This means that $(f_1^*(x'), \ldots, f_m^*(x'))$ is pure-strategy equilibrium with $\mathcal{D}_{x'}$. We characterize the equilibrium structure for $\mathcal{D}_{x'}$ and use this characterization to compute the equilibrium social loss.

Equilibrium structure for $\mathcal{D}_{x'}$. For notational convenience, let $y_i := f_i^*(x')$ denote the label chosen by model-provider i and let let $y^* = \arg\max_y \mathbb{P}[y \mid x']$ be the Bayes optimal label for x'. We also abuse notation slightly and let $u(y_1; y_{-j})$ be model-provider 1's utility if they choose the label y_1 for x' and the other model-provider's choose y_{-j}.

We first show that all model-providers choosing y^* is an equilibrium if and only if $\alpha(x') \le 1/m$. Let's fix $y_j = y^*$ for all $j \ge 2$ and look at model-provider 1's utility. We see that $u(y^*; y_{-j}) = 1/m$ and $u(1 - y^*; y_{-j}) = \alpha(x')$. This means that y^* is a best-response (i.e., $y^* \in \arg\max_y u(y; y_{-j})$) if and only if $\alpha(x') \le 1/m$.

We next show that if $\alpha(x') < 1/m$, then the market outcome $y_i = y^*$ for all $i \in [m]$ is the only pure-strategy equilibrium. Let $y_1, \ldots, y_m$ be a pure-strategy equilibrium. It suffices to show that y^* is the unique best response to y_{-j}; that is, that $\{y^*\} = \arg\max_y u(y; y_{-j})$. To show this, let m' denote the size of the set $\{2 \le i \le m \mid y_i = y^*\}$. First, if $m' = 0$, then we have that

$$u(y^*; y_{-j}) = 1 - \alpha(x') > 1/m = u(1 - y^*; y_{-j}),$$

where $1-\alpha(x') > 1/m$ follows from the fact that $1-\alpha(x') \geq 1/2 \geq 1/m$ along with our assumption that $\alpha(x') \neq 1/m$. This demonstrates that y^* is indeed the unique best response. If $m' = m-1$, then we have that:

$$u(y^*; y_{-j}) = 1/m > \alpha(x') = u(1-y^*; y_{-j}),$$

as desired. Finally, if $1 \leq m' \leq m-2$, then:

$$u(y^*; y_{-j}) = \frac{1-\alpha(x')}{m'+1} \geq \frac{1-\alpha(x')}{m-1} > \frac{1}{m} > \alpha(x') > \frac{\alpha(x')}{m-m'} = u(1-y^*; y_{-j}),$$

as desired.

Finally, we show that all model-providers choosing $1-y^*$ is never an equilibrium. Let's fix $y_j = 1-y^*$ and look at model-provider 1's utility. We see that:

$$u(y^*; y_{-j}) = 1-\alpha(x') > \frac{\alpha(x')}{m} = u(1-y^*; y_{-j}),$$

which shows that y^* is the unique best response as desired.

Characterization of equilibrium social loss. It follows from (5) that the equilibrium social loss $\mathbb{E}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f^*_{j^*(x,y)}(x), y)]$ is $\alpha(x')$ if all of the model-providers choose $y_i = y^*$, it is zero if a nonzero number of model-providers choose y^* and a nonzero number of model-providers choose $1-y^*$, and it is $1-\alpha(x')$ if all of the model-providers choose $1-y^*$.

Let's combine this with our equilibrium characterization results. If $\alpha(x') < 1/m$, then the unique equilibrium is at $y_i = y^*$ so the equilibrium social loss is $\alpha(x)$ as desired. If $\alpha(x') > 1/m$, then neither $y_i = y^*$ for all $i \in [m]$ nor $y_i = 1-y^*$ for all $i \in [m]$ is an equilibrium. Since there exists a pure strategy equilibrium by Proposition 1, there must be a pure strategy equilibrium where a nonzero number of model-providers choose y^* and a nonzero number of model-providers choose $1-y^*$. The equilibrium social loss is thus zero.

Note when $\alpha(x') = 1-1/m$, there is actually an equilibrium where all of the model-providers choose $y_i = y^*$, 0 and an equilibrium where a nonzero number of model-providers choose y^* and a nonzero number of model-providers choose $1-y^*$; thus, the equilibrium social loss can be zero or $1/m$.

□

B.3 Proofs for Section 3.4

We prove Proposition 3.

Proof of Proposition 3. Let $\mathcal{D}^R$ be the marginal distribution of X with respect to the distribution $(X,Y) \sim \mathcal{D}$. Let $f^*_1, \ldots, f^*_m$ be a pure-strategy equilibrium. The social loss is equal to:

$$\begin{aligned}
\texttt{SL}(f_1, \ldots, f_m) &= \mathbb{E}[\ell(f^*_{j^*(x,y)}(x), y)] \\
&= \mathop{\mathbb{E}}_{x'\sim\mathcal{D}^R}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(f^*_{j^*(x,y)}(x), y) \mid x = x']\right] \\
&= \mathop{\mathbb{E}}_{x'\sim\mathcal{D}^R}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f^*_{j^*(x,y)}(x), y)]\right],
\end{aligned}$$

where $\mathcal{D}_{x'}$ denotes the conditional distribution $(X, Y) \mid X = x'$ where $(X, Y) \sim \mathcal{D}$. Thus, to analyze the overall social loss, we can separately analyze the social loss on each distribution $\mathcal{D}_{x'}$ and then average across distributions. It suffices to show that

$$\mathop{\mathbb{E}}_{\mathcal{D}_{x'}}\left[\sum_{i=1}^{K}\alpha^i(x)\cdot\mathbb{1}\left[\alpha^i(x)<\frac{c}{m}\right]\right]\le\mathop{\mathbb{E}}_{\mathcal{D}_{x'}}\left[\ell(f^*_{j^*(x,y)}(x),y)\right]\le\mathop{\mathbb{E}}_{\mathcal{D}_{x'}}\left[\sum_{i=1}^{K}\alpha^i(x)\cdot\mathbb{1}\left[\alpha^i(x)\le\frac{1}{m}\right]\right].$$

To compute the social loss on $\mathcal{D}_{x'}$, we first apply Lemma 1. This means that $(f^*_1(x'),\dots,f^*_m(x'))$ is pure-strategy equilibrium with $\mathcal{D}_{x'}$. We then prove properties of the equilibrium structure for $\mathcal{D}_{x'}$ and use these properties to bound the equilibrium social loss. For notational convenience, let $y_i := f^*_i(x')$ denote the label chosen by model-provider i and let let $y^* = \arg\max_y \mathbb{P}[y \mid x']$ be the Bayes optimal label for x'. We also abuse notation slightly and let $u(y_1; y_{-j})$ be model-provider 1's utility if they choose the label y_1 for x' and the other model-provider's choose y_{-j}. We can rewrite:

$$\mathop{\mathbb{E}}_{\mathcal{D}_{x'}}\left[\ell(f^*_{j^*(x,y)}(x),y)\right]=\mathop{\mathbb{E}}_{\mathcal{D}_{x'}}\left[\sum_{i=1}^{K}\alpha^i(x)\cdot\mathbb{1}\left[y_j\neq i\text{ for all }j\in[m]\right]\right].$$

We first prove the lower bound on $\mathbb{E}_{\mathcal{D}_{x'}}[\ell(f^*_{j^*(x,y)}(x),y)]$ and then we prove the upper bound on $\mathbb{E}_{\mathcal{D}_{x'}}[\ell(f^*_{j^*(x,y)}(x),y)]$.

Proof of lower bound. Let $y_1,\dots,y_m$ be a pure strategy equilibrium. To prove the lower bound, it suffices to show that if $\alpha^i(x) < c/m$, then $y_j \neq i$ for all $j \in [m]$.

Assume for sake of contradiction that $\alpha^i(x) < c/m$ and $y_j = i$ for some $j \in [m]$. Let $i' = \text{argmax}_{i''\in\{0,1,\dots,K-1\}}\alpha^{i''}(x)$ be the class with maximal conditional probability. By the definition of c, we see that $\alpha^{i'}(x) \geq c > c/m$ which also implies that $i' \neq i$. We split into two cases—(1) $y_{j'} \neq i'$ for all $j' \in \{0, 1, \dots, K-1\}$, and (2) $y_{j'} = i'$ for some $j' \in \{0, 1, \dots, K-1\}$—and derive a contradiction in each case.

Consider the first case where $y_{j'} \neq i'$ for all $j' \in \{0, 1, \dots, K-1\}$. Then if model-provider j switched from y_j to i', the difference in their utility would be bounded as:

$$\begin{aligned}
u(i'; y_{-j}) - u(y_j; y_{-j}) &\geq \alpha^{i'}(x) - \left(\frac{\alpha^{i'}(x)}{m} + \alpha^i(x)\right)\\
&= \alpha^{i'}(x)\left(1-\frac{1}{m}\right) - \alpha^i(x)\\
&> c\left(1-\frac{1}{m}\right) - \frac{c}{m}\\
&= c\left(1-\frac{2}{m}\right)\\
&\geq 0,
\end{aligned}$$

so y_j is not a best-response for model-provider j, which is a contradiction.

Now, consider the second case, where $y_{j'} = i'$ for some $j' \in \{0, 1, \dots, K-1\}$. If we compare the utility when model-provider j chooses i' versus y_j as their action, the difference is utility can be bounded as:

$$u(i'; y_{-j}) - u(y_j; y_{-j}) \geq \frac{\alpha^{i'}(x)}{m} - \alpha^i(x) > \frac{c}{m} - \frac{c}{m} = 0.$$

so y_j is not a best-response for model-provider j, which is a contradiction.

This proves the lower bound as desired.

	$y_2 = 1 - y^*$	$y_2 = y^*$
$y_1 = 1 - y^*$	$(w_{\max}, w_{\min})$	$(\alpha(x), 1 - \alpha(x))$
$y_1 = y^*$	$(1 - \alpha(x), \alpha(x))$	$(w_{\max}, w_{\min})$

Table 1: Let $X = \{x\}$, $\mathcal{F} = \mathcal{F}_{\text{all}}^{\text{binary}}$, user decisions are noiseless, and user decisions are noiseless (i.e., $c \to 0$, so user decisions are given by (8)). Suppose that there are $m = 2$ model-providers with market reputations $w_{\min}$ and $w_{\max}$, where $w_{\max} \geq w_{\min}$ and $w_{\max} + w_{\min} = 1$. Let $y^* = \arg\max_y \mathbb{P}[y \mid x]$ be the Bayes optimal label for x'. The table shows the game matrix when model-provider 1 chooses the label y_1 and model provider 2 chooses the label y_2.

Proof of upper bound. Let $y_1, \ldots, y_m$ be a pure strategy equilibrium. To prove the upper bound, it suffices to show if $\alpha^i(x) > 1/m$, then $y_j = i$ for some $j \in [m]$. Assume for sake of contradiction that $\alpha^i(x) > 1/m$ and $y_j \neq i$ for all $j \in [m]$. For any set of actions $y_1, \ldots, y_m$, the total utility $\sum_{j=1}^m u(y_j; y_{-j}) = 1$ sums to 1. Thus, some model provider $j \in [m]$ must have utility satisfying $u(y_j; y_{-j}) \leq 1/m$. However, if model-provider j instead chose action i, then they would achieve utility:

$$u(i; y_{-j}) \geq \alpha^i(x) > \frac{1}{m} \geq u(y_j; y_{-j}),$$

so y_j is not a best-response for model-provider j, which is a contradiction. This proves the upper bound as desired.

□

B.4 Proofs for Section 3.5

A useful lemma is the following calculation of the game matrix when there is a single representation $X = \{x\}$.

Lemma 2. *Let $X = \{x\}$, and let $\mathcal{F} = \mathcal{F}_{all}^{binary}$. Suppose that there are $m = 2$ model-providers with market reputations w_{min} and w_{max}, where $w_{max} \geq w_{min}$ and $w_{max} + w_{min} = 1$. Suppose that user decisions are noiseless (i.e., $c \to 0$, so user decisions are given by (8)). Then the game matrix is specified by Table 1.*

Proof. This follows from applying (8) and using the fact that $\ell(y, y') = \mathbb{1}[y \neq y']$. □

We show that pure strategy equilibria are no longer guaranteed to exist when model-providers have unequal market reputations, even when there is a single representation $X = \{x\}$.

Lemma 3. *Let $X = \{x\}$ let $\mathcal{F} = \mathcal{F}_{all}^{binary}$. Suppose that there are $m = 2$ model-providers with market reputations w_{min} and w_{max}, where $w_{max} \geq w_{min}$ and $w_{max} + w_{min} = 1$. Suppose that user decisions are noiseless (i.e., $c \to 0$, so user decisions are given by (8)). If $\alpha(x) > w_{min}$, then a pure strategy equilibrium does not exist.*

Proof. For notational convenience, let $y_i := f_i(x')$ denote the label chosen by model-provider i and let $y^* = \arg\max_y \mathbb{P}[y \mid x']$ be the Bayes optimal label for x'. We also abuse notation slightly and let $u_i(y; y')$ be model-provider i's utility if they choose the label y for x and the other model-providers choose y'. The proof follows from the game matrix show in Table 1 (Lemma 2). Using the fact that model-provider 1 must best-respond to model-provider 2's action, this leaves $y_1 = 1 - y^*$,

$y_2 = 1 - y^*$ and $y_1 = y^*$, $y_2 = y^*$. However, neither of these market outcomes captures a best-response for model-provider 2: if $y_1 = 1 - y^*$, then model-provider 2's unique best response is y^*; if $y_1 = y^*$, then model-provider 2's unique best response is $1 - y^*$. This rules out the existence of a symmetric or asymmetric pure strategy equilibrium. □

Given the lack of existence of pure strategy equilibria, we must turn to mixed strategies. A mixed strategy equilibrium is guaranteed to exist since the game has finitely many actions $\mathcal{F}_{\text{all}}^{\text{binary}}$ and finitely many players m. Let $(\mu_1, \mu_2, \ldots, \mu_m)$ denote a mixed strategy profile over $\mathcal{F}_{\text{all}}^{\text{binary}}$. We show the following analogue of Lemma 1 that allows us to again decompose model-provider actions into independent decisions about each representation x. To formalize this, let $\mathcal{D}$ be the data distribution, and again let $\mathcal{D}_x$ be the conditional distribution of (X, Y) when $X = x$, where $(X, Y) \sim \mathcal{D}$. Again, let $(\mathcal{F}_{\text{all}}^{\text{binary}})^x := \{f_0, f_1\}$ be the class of the (two) functions from a single representation x to $\{0, 1\}$, where $f_0(x) = 0$ and $f_1(x) = 1$. Given a mixed strategy profile μ and a representation x, we define the conditional mixed strategy μ^x over $(\mathcal{F}_{\text{all}}^{\text{binary}})^x := \{f_0, f_1\}$ to be defined so $\mathbb{P}_{\mu^x}[f_i] := \mathbb{P}_{f\sim\mu}[f(x) = i]$ for $i \in \{0, 1\}$.

Lemma 4. *Let X be a finite set of representations, let $\mathcal{F} = (\mathcal{F}_{all}^{binary})$, and let $\mathcal{D}$ be the distribution over (X, Y). For each $x \in X$, let $\mathcal{D}_x$ be the conditional distribution of (X, Y) given $X = x$, where $(X, Y) \sim \mathcal{D}$, and let $(\mathcal{F}_{all}^{binary})^x := \{f_0, f_1\}$ be the class of the (two) functions from a single representation x to $\{0, 1\}$, where $f_0(x) = 0$ and $f_1(x) = 1$. Suppose that user decisions are noiseless (i.e., $c \to 0$, so user decisions are given by (3)). A strategy profile $(\mu_1, \mu_2, \ldots, \mu_m)$ is an equilibrium if and only if for every $x \in X$, the market outcome $(\mu_1^x, \mu_2^x, \ldots, \mu_m^x)$ (where μ_1^x, …, μ_m^x are the conditional mixed strategies defined above) is an equilibrium for $(\mathcal{F}_{all}^{binary})^x$ with data distribution $\mathcal{D}_x$.*

Proof. The proof follows similarly to the proof of Lemma 4, but some minor generalizations to account for mixed strategy equilibria. Let $\mathcal{D}^R$ be the marginal distribution of X with respect to the distribution $(X, Y) \sim \mathcal{D}$. Let $\mathcal{D}^R$ be the marginal distribution of X with respect to the distribution $(X, Y) \sim \mathcal{D}$. First, we write model-provider j's utility as:

$$\mathop{\mathbb{E}}_{\substack{f_j\sim\mu_j \\ \mathbf{f}_{-j}\sim\mu_{-j}}}[u(f_j; \mathbf{f}_{-j})] = \mathop{\mathbb{E}}_{\substack{f_j\sim\mu_j \\ \mathbf{f}_{-j}\sim\mu_{-j}}}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\mathbb{P}[j^*(x, y) = j]]\right] = \mathop{\mathbb{E}}_{x'\sim\mathcal{D}^R}\left[\mathop{\mathbb{E}}_{\substack{f_j\sim\mu_j^{x'} \\ \mathbf{f}_{-j}\sim\mu_{-j}^{x'}}}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\mathbb{P}[j^*(x, y) = j]]\right]\right]. \tag{12}$$

where μ_{-j} denotes the mixed strategies chosen by the other model-providers.

First we show that if $\mu_1, \mu_2, \ldots, \mu_m$ is an equilibrium, then $(\mu_1^{x'}, \ldots, \mu_m^{x'})$ is an equilibrium for $(\mathcal{F}_{\text{all}}^{\text{binary}})^{x'}$ with data distribution $\mathcal{D}_{x'}$. Let f_j be in $\text{supp}(\mu_{j'})$. Assume for sake of contradiction that $(\mu_1^{x'}, \ldots, \mu_m^{x'})$ is not an equilibrium. Then there exists $j' \in [m]$ such that model-provider j' would achieve higher utility on $f^{1-f_{j'}(x')}$ than $f^{f_{j'}(x')}$. Let $f'_{j'}$ be the predictor given by $f'_{j'}(x) = f_{j'}(x)$ if $x \neq x'$ and $f'_{j'}(x') = 1 - f_{j'}(x')$. By equation (12), this would mean that $u(f'_{j'}; \mu_{-j'})$ is strictly higher than $u(f_{j'}; \mu_{-j'})$ which is a contradiction.

Next, we show that if $(\mu_1^{x'}, \ldots, \mu_m^{x'})$ is an equilibrium for $(\mathcal{F}_{\text{all}}^{\text{binary}})^{x'}$ with data distribution $\mathcal{D}_{x'}$ for all $x' \in X$ then $\mu_1, \ldots, \mu_m$ is an equilibrium. Let f_j be in $\text{supp}(\mu_{j'})$. Assume for sake of contradiction that there exists j' such that $u(f'_{j'}; \mu_{-j'}) > u(f_j; \mu_{-j'})$. By equation (11), there must exist x' such that $\mathbb{E}_{\mathbf{f}_{-j'}\sim\mu_{-j'}^{x'}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_{x'}}[\mathbb{P}[j^*(x, y) = j']]\right]$ is higher for $f'_{j'}$ than $f_{j'}$. This means that $(\mu_1^{x'}, \ldots, \mu_m^{x'})$ is not an equilibrium, which is a contradiction. □

We now prove Proposition 4.

Proof of Proposition 4. Let $\mathcal{D}^R$ be the marginal distribution of x with respect to the distribution $(x, y) \sim \mathcal{D}$. Let μ_1, μ_2 be a mixed strategy equilibrium. The social loss is equal to:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1\\ f_2\sim\mu_2}}[\mathtt{SL}(f_1, f_2)] &= \mathbb{E}[\ell(f_{j^*(x,y)}(x), y)] \\
&= \mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1\\ f_2\sim\mu_2}}\left[\mathop{\mathbb{E}}_{x'\sim\mathcal{D}^R}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(f_{j^*(x,y)}(x), y) \mid x = x']\right]\right] \\
&= \mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1\\ f_2\sim\mu_2}}\left[\mathop{\mathbb{E}}_{x'\sim\mathcal{D}^R}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f_{j^*(x,y)}(x), y)]\right]\right] \\
&= \mathop{\mathbb{E}}_{x'\sim\mathcal{D}^R}\left[\mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1^*\\ f_2\sim\mu_2^*}}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f_{j^*(x,y)}(x), y)]\right]\right] \\
&= \mathop{\mathbb{E}}_{x'\sim\mathcal{D}_X}\left[\mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1^{x'}\\ f_2\sim\mu_2^{x'}}}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f_{j^*(x,y)}(x), y)]\right]\right]
\end{aligned}$$

where $\mathcal{D}_{x'}$ denotes the conditional distribution $(X, Y) \mid X = x'$ where $(X, Y) \sim \mathcal{D}$ and where μ^x denotes the conditional mixed strategy $(\mathcal{F}_{\text{all}}^{\text{binary}})^x := \{f^0, f^1\}$ to be defined so $\mathbb{P}_{\mu^x}[f^i] := \mathbb{P}_{f\sim\mu}[f(x) = i]$ for $i \in \{0, 1\}$ Thus, to analyze the overall social loss, we can separately analyze the social loss on each distribution $\mathcal{D}_{x'}$ and then average across distributions. It suffices to show that:

$$\mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1^{x'}\\ f_2\sim\mu_2^{x'}}}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f_{j^*(x,y)}(x), y)]\right] = \begin{cases} \alpha(x') & \text{if } \alpha(x') < w_{\min} \\ \frac{2(\alpha(x') - w_{\min})\cdot(w_{\max} - \alpha(x))}{(1-2\cdot w_{\min})^2} & \text{if } \alpha(x') > w_{\min}. \end{cases}$$

To compute the social loss on $\mathcal{D}_{x'}$, we first apply Lemma 4. This means that $(\mu_1^{x'}, \mu_2^{x'})$ is mixed-strategy equilibrium with $\mathcal{D}_{x'}$. We characterize the equilibrium structure for $\mathcal{D}_{x'}$ and use this characterization to compute the equilibrium social loss.

Our main technical ingredient is the game matrix in Table 1 (Lemma 2). We will slightly abuse notation and view choosing the label y as the strategy of the model-provider. Accordingly, we view a mixed strategy as a distribution over $\{0, 1\}$. For notational convenience, let $y_i := f_i(x')$ denote the label chosen by model-provider i and let $y^* = \arg\max_y \mathbb{P}[y \mid x']$ be the Bayes optimal label for x'. We split into two cases: $\alpha(x') < w_{\min}$ and $\alpha(x') > w_{\min}$.

Case 1: $\alpha(x') < w_{\min}$. We claim that the unique equilibrium is a pure strategy equilibrium where $y_1 = y_2 = y^*$. First, if $\alpha(x) < w_{\min}$, we show that choosing y^* is a strictly dominant strategy for model-provider 1. This follows from the fact that $1 - \alpha(x) > w_{\max}$ and $w_{\max} \geq w_{\min} > \alpha(x)$. Thus, model-provider 1 must play a pure strategy where they always choose $y_1 = y^*$. When model-provider 1 chooses y^*, then the unique best response for model-provider 2 is also to choose y^* since $\alpha(x') < w_{\min}$. This establishes that $y_1 = y_2 = y^*$ is the unique equilibrium. This also implies that

the equilibrium social loss satisfies:

$$\mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1^{x'}\\ f_2\sim\mu_2^{x'}}}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f_{j^*(x,y)}(x),y)]\right]=\alpha(x')$$

as desired.

Case 2: $\alpha(x') > w_{\min}$**.** Let $p_1 = \mathbb{P}_{\mu_1^{x'}}[y_1 = y^*]$ and let $p_2 = \mathbb{P}_{\mu_2^{x'}}[y_2 = y^*]$. By Lemma 3, a pure strategy equilibrium does not exist. Thus, we consider mixed strategies. Since pure strategy equilibria do not exist, at least one of p_1 and p_2 must be strictly between zero and one. We compute p_1 and p_2, splitting into two cases: (1) $p_1 > 0$ and (2) $p_2 > 0$.

If $p_1 > 0$, then we know that model-provider 1 must be indifferent between choosing y^* and $1 - y^*$. This means that:

$$p_2\alpha(x') + (1-p_2)w_{\max} = (1-p_2)(1-\alpha(x')) + p_2 w_{\max}.$$

Solving for p_2, we obtain:

$$p_2 = \frac{w_{\max} - (1-\alpha(x'))}{2w_{\max}-1} = \frac{\alpha(x') - w_{\min}}{1-2w_{\min}} > 0.$$

If $p_2 > 0$, then we know that model-provider 2 must be indifferent between choosing y^* and $1 - y^*$. This means that:

$$p_1\alpha(x') + (1-p_1)w_{\min} = (1-p_1)(1-\alpha(x')) + p_1 w_{\min}.$$

Solving for p_1, we obtain:

$$p_1 = \frac{(1-\alpha(x')) - w_{\min}}{1-2w_{\min}} = \frac{w_{\max} - \alpha(x)}{1-2w_{\min}} > 0.$$

Putting this all together, we see that:

$$p_1 = \frac{w_{\max} - \alpha(x')}{1-2w_{\min}}$$
$$p_2 = \frac{\alpha(x') - w_{\min}}{1-2w_{\min}},$$

and in fact $p_1 + p_2 = 1$.

Using this characterization of p_1 and p_2, we see that the equilibrium social loss is equal to:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{f_1\sim\mu_1^{x'}\\ f_2\sim\mu_2^{x'}}}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{x'}}[\ell(f_{j^*(x,y)}(x),y)]\right] &= \alpha(x')\mathbb{P}[y_1 = y^*]\mathbb{P}[y_2 = y^*] + (1-\alpha(x'))\mathbb{P}[y_1 = 1-y^*]\mathbb{P}[y_2 = 1-y^*] \\
&= \alpha(x')p_1p_2 + (1-\alpha(x))(1-p_1)(1-p_2) \\
&= \alpha(x')p_1p_2 + (1-\alpha(x))p_1p_2 \\
&= p_1p_2 \\
&= \frac{(\alpha(x') - w_{\min})\cdot(w_{\max} - \alpha(x))}{(1-2\cdot w_{\min})^2},
\end{aligned}$$

as desired.

□