Incentivizing honest performative predictions with proper scoring rules

Caspar Oesterheld*[1] Johannes Treutlein*[2] Emery Cooper[3] Rubi Hudson[4]

[1]Carnegie Mellon University
[2]University of California, Berkeley
[3]Center on Long-Term Risk
[4]University of Toronto

Abstract

Proper scoring rules incentivize experts to accurately report beliefs, assuming predictions cannot influence outcomes. We relax this assumption and investigate incentives when predictions are *performative*, i.e., when they can influence the outcome of the prediction, such as when making public predictions about the stock market. We say a prediction is a *fixed point* if it accurately reflects the expert's beliefs after that prediction has been made. We show that in this setting, reports maximizing expected score generally do not reflect an expert's beliefs, and we give bounds on the inaccuracy of such reports. We show that, for binary predictions, if the influence of the expert's prediction on outcomes is bounded, it is possible to define scoring rules under which optimal reports are arbitrarily close to fixed points. However, this is impossible for predictions over more than two outcomes. We also perform numerical simulations in a toy setting, showing that our bounds are tight in some situations and that prediction error is often substantial (greater than 5-10%). Lastly, we discuss alternative notions of optimality, including performative stability, and show that they incentivize reporting fixed points.

1 INTRODUCTION

As AI capabilities increase, this raises concern for safety, including how to scalably control AI systems with superhuman capabilities [Russell, 2019, Ngo et al., 2022]. One proposed design for safety is *oracle AI* [Armstrong et al., 2012; Armstrong, 2013; Bostrom, 2014, Ch. 10]. An oracle AI makes predictions or forecasts about the world, but does not autonomously pursue goals. It could thus be safer while still being useful for many applications.

A proper scoring rule assigns scores to forecasts in a way that incentivizes honest reporting of beliefs [Brier, 1950; Good, 1952, Section 8; McCarthy, 1956; Savage, 1971; Gneiting and Raftery, 2007]. Proper scoring rules have been used to incentivize honest reports from experts [Carvalho, 2016]. They could thus be used as an objective for oracle AIs. However, prior work assumes that predictions themselves do not influence the events they are trying to predict. In reality, predictions may be *performative* [Perdomo et al., 2020, Armstrong and O'Rorke, 2017], meaning that they can influence the distribution of outcomes. For example, an AI predicting stock market prices might be able to influence whether people buy or sell stocks, and thus influence whether its predictions come true or not. This makes it important to investigate incentives and honesty of predictions when predictions are performative.

In this paper, we analyze the case of an AI model or human, henceforth called expert, making a probabilistic forecast over a finite set of possibilities to maximize a proper scoring rule. We say that a prediction is performatively optimal if it maximizes expected score, and we define a prediction as a *fixed point* or self-fulfilling if it is equal to the expert's beliefs, conditional on the expert having made that prediction. We investigate to what extent honest predictions, i.e., fixed points, are incentivized in this setting.[1] All else equal, honest predictions are preferable since, assuming a sufficiently capable expert, they provide us with more accurate information. However, if an expert has incentives other than to predict honestly—e.g., to bring about fixed points with lower entropy—this is undesirable even if the expert otherwise makes approximately accurate predictions.

The setting in which a model's predictions can influence the predicted distribution has been discussed as *performative prediction* [Perdomo et al., 2020] in the machine learning

*Equal contribution

[1] We assume that the AI model can be ascribed explicit beliefs, so that its reports can be characterized as honest if they reflect the model's beliefs.

literature. However, performative prediction focuses on classification or regression tasks with arbitrary model classes and loss functions rather than probabilistic predictions incentivized by proper scoring rules. The literature is motivated by minimization of a given loss function, whereas we take a mechanism design perspective, asking which scoring rules incentivize honest predictions. Focusing on a special case and taking a different perspective will lead to original results that are unique to our setting.

Contributions. In Section 3, we adapt the performative prediction formalism to probabilistic predictions or forecasts. We allow for an arbitrary function f describing the relationship between the expert's predictions and distributions over predicted outcomes caused by these predictions.

In Section 4, we show that for any strictly proper scoring rule, **there exist functions f from predictions to beliefs such that performatively optimal reports are not fixed points**, even if one exists and is unique. Moreover, we show that under reasonable distributions over such functions, **optimal reports are almost never fixed points**. This strengthens analogous results from the performative prediction literature.

In Section 5, we then **provide upper bounds** for the inaccuracy of reported beliefs, and for the distance of predictions from fixed points.

In Section 6, we use the bounds to develop **scoring rules that make the bounds arbitrarily small for binary predictions**. We also show that **when reporting a prediction over more than two outcomes, the bounds cannot be made arbitrarily small**.

In Section 7, we perform **numerical simulations using the quadratic scoring rule**, to show how the inaccuracy of predictions and the distance of predictions from fixed points depend on the expert's influence on the world via its prediction. The results show that our bounds are tight in some cases. They also show that substantially inaccurate reports (i.e., with errors greater than $5 - 10\%$) are common in our toy setting.

In Section 8, we discuss alternatives to performative optimality that do not set incentives other than honest predictions. We show that *performatively stable* [Perdomo et al., 2020] predictions are fixed points. We then consider repeated risk minimization, repeated gradient descent, no-regret learning and prediction markets, and show that all of these settings lead to predictions that are fixed points or close to fixed points.

Finally, in Section 9, we elaborate on related work, and in Section 10, we conclude and outline avenues for future work.

Proofs are in corresponding sections in Appendix A.

2 BACKGROUND

Proper scoring rules. Proper scoring rules are used to incentivize an expert to report probabilistic beliefs honestly. Consider a prediction given by a probability distribution $\boldsymbol{p} \in \Delta(\mathcal{N})$ over a set $\mathcal{N} := \{1, \ldots, n\}$ of $n \in \mathbb{N}$ disjoint and exhaustive outcomes. We identify each distribution $\boldsymbol{p} \in \Delta(\mathcal{N})$ with a vector $\boldsymbol{p} \in [0,1]^n$ and write p_i for the probability of event $i \in \mathcal{N}$ under distribution $\boldsymbol{p}$. A *scoring rule* is a function $S\colon \Delta(\mathcal{N}) \times \mathcal{N} \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} := [-\infty, \infty]$ is the extended real line. Given prediction $\boldsymbol{p} \in \Delta(\mathcal{N})$ and outcome $i \in \mathcal{N}$, the expert receives the score $S(\boldsymbol{p}, i)$. We write $S(\boldsymbol{p}, \boldsymbol{q}) := \mathbb{E}_{i \sim \boldsymbol{q}}[S(\boldsymbol{p}, i)]$ for the expert's expected score, given that outcome i follows distribution $\boldsymbol{q} \in \Delta(\mathcal{N})$.

Definition 1. A scoring rule S is called *proper* if $S(\boldsymbol{q}, \boldsymbol{q}) \geq S(\boldsymbol{p}, \boldsymbol{q})$ for all $\boldsymbol{p}, \boldsymbol{q} \in \Delta(\mathcal{N})$. It is called *strictly proper* if this inequality is strict whenever $\boldsymbol{p} \neq \boldsymbol{q}$.

Example 1 (Logarithmic scoring rule)**.** The logarithmic scoring rule is defined as $S(\boldsymbol{p}, i) := \log p_i$ and $S(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{n} q_i \log p_i$. This is also the negative of the cross-entropy loss employed in training, for example, current large language models [Brown et al., 2020]. It is strictly proper.

Example 2 (Quadratic scoring rule)**.** Another strictly proper scoring rule is the quadratic score, defined as $S(\boldsymbol{p}, i) := 2p_i - \|\boldsymbol{p}\|_2^2$ with $S(\boldsymbol{p}, \boldsymbol{q}) = 2\boldsymbol{p}^\top \boldsymbol{q} - \|\boldsymbol{p}\|_2^2$. This is an affine transformation of the Brier score, making them equivalent scoring rules.

Gneiting and Raftery [2007, Theorem 1] provide a characterization of proper scoring rules, which will be helpful for stating and proving many of our results.

First, given a convex function $G\colon \Delta(\mathcal{N}) \to \overline{\mathbb{R}}$, a subgradient is a function $g\colon \Delta(\mathcal{N}) \to \overline{\mathbb{R}}^n$ such that for any $\boldsymbol{p}, \boldsymbol{q} \in \Delta(\mathcal{N})$, we have $G(\boldsymbol{q}) \geq G(\boldsymbol{p}) + g(\boldsymbol{p})^\top (\boldsymbol{q} - \boldsymbol{p})$. In general, this function may not be unique. Throughout this paper we assume that whenever the subgradients are finite, they are normalized to lie in the *tangent space* of $\Delta(\mathcal{N})$, i.e., $g(\boldsymbol{p}) \in \mathcal{T} := \{\boldsymbol{x} \in \mathbb{R}^n \mid \sum_i x_i = 0\}$. This can be assumed since if $g(\boldsymbol{p})$ is a subgradient of G at point $\boldsymbol{p}$, so is $(g_i(\boldsymbol{p}) - \frac{1}{n}\sum_j g_j(\boldsymbol{p}))_i$.

Theorem 1 (Gneiting and Raftery, 2007)**.** *A scoring rule S is (strictly) proper, if and only if there exists a (strictly) convex function $G\colon \Delta(\mathcal{N}) \to \overline{\mathbb{R}}$ with a* subgradient *$g\colon \Delta(\mathcal{N}) \to \overline{\mathbb{R}}^n$ such that $S(\boldsymbol{p}, \boldsymbol{q}) = G(\boldsymbol{p}) + g(\boldsymbol{p})^\top (\boldsymbol{q} - \boldsymbol{p})$ for all $\boldsymbol{p}, \boldsymbol{q} \in \Delta(\mathcal{N})$.*

Differentiable scoring functions. If G is differentiable at some point $\boldsymbol{p}$, then the subgradient $g(\boldsymbol{p})$ is just the gradient of G, $g(\boldsymbol{p}) = \nabla G(\boldsymbol{p})$. As before we let $\nabla G(\boldsymbol{p})$ be an element of the tangent space $\mathcal{T}$. For any $\boldsymbol{v} \in \mathcal{T}$, $g(\boldsymbol{p})^\top \boldsymbol{v}$ then gives the directional derivative of G at point $\boldsymbol{p}$ in the

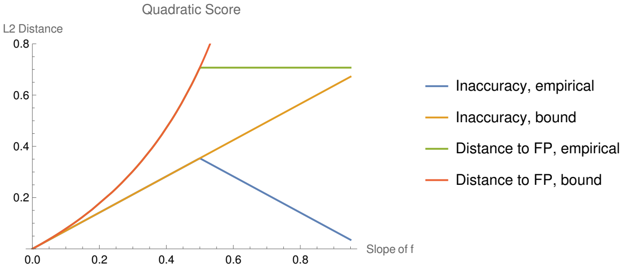

Figure 1: Maximal inaccuracy and maximal distance to fixed point (FP) of optimal predictions, depending on the slope of f, according to our simulation and our theoretical bound.

direction $\boldsymbol{v}$. Note that since G is only defined on the simplex $\Delta(\mathcal{N})$, the partial derivatives are not well-defined.[2]

Given G, g as in the Gneiting and Raftery characterization, we write $Dg(\boldsymbol{p}) \in \mathbb{R}^{n,n}$ for the Jacobian matrix of g, if it exists (i.e., this is the Hessian of G). Note that because g is only defined on $\Delta(\mathcal{N})$, the matrix representation of $Dg(\boldsymbol{p})$ in $\mathbb{R}^{n,n}$ is not unique. Generally it does not matter which representation of $Dg(\boldsymbol{p})$ we use. Importantly, for all $\boldsymbol{v} \in \mathcal{T}$, $Dg(\boldsymbol{p})\boldsymbol{v}$ will always be unique and (because we assume that $g(\boldsymbol{p})$ is in the tangent space) in $\mathcal{T}$.

3 PROBLEM SETTING

In this paper, we take the stance of a principal trying to elicit honest predictions from an expert (human or AI system). We assume that the expert reports a prediction $\boldsymbol{p}$ to maximize the expected score given by a proper scoring rule, $S(\boldsymbol{p}, \boldsymbol{q})$.

Importantly, we assume that the expert's beliefs over outcomes, $\boldsymbol{q}$, can themselves vary given different predictions $\boldsymbol{p}$, because the expert may believe that its predictions affect the probability of outcomes. To model this, we assume that there is a function $f \colon \Delta(\mathcal{N}) \to \Delta(\mathcal{N})$ such that beliefs are given by $\boldsymbol{q} = f(\boldsymbol{p})$.[3] We assume f is known to the expert, but not to the principal.

In the case of an AI system, $f(\boldsymbol{p})$ could also be seen as a ground distribution from which we sample to *train* a model (see Appendix D.3). In that case, the objective is to design a training procedure that sets the right incentives. However, in most of the following, we assume $f(\boldsymbol{p})$ are the subjective beliefs of a highly capable and knowledgeable expert.

We say that a prediction $\boldsymbol{p}$ is *performatively optimal* [Perdomo et al., 2020] if $\boldsymbol{p} \in \arg\max_{\Delta(\mathcal{N})} S(\boldsymbol{p}, f(\boldsymbol{p}))$. In the following, we will not assume convexity of this objective. Our bounds will depend on differentiability of S and f.

A point $\boldsymbol{p}$ is a *fixed point* of f if $f(\boldsymbol{p}) = \boldsymbol{p}$. By Brouwer's fixed point theorem, if f is continuous, a fixed point $\boldsymbol{p} \in \Delta(\mathcal{N})$ always exists. Moreover, if f is Lipschitz continuous with constant $L_f < 1$, then by Banach's fixed point theorem, the fixed point is unique.

Example 3 (Bank Run)**.** A newspaper's AI predicts whether a certain bank will suffer a bank run or not. Readers use this information when deciding whether to withdraw their money. Specifically, imagine that the probability of a bank run as a function of the AI expert's prediction $\mathbf{p} = (p_1, p_2) \in \Delta(\{1, 2\})$ is given by the (monotonic) function $f \colon \Delta(\{1, 2\}) \to \Delta(\{1, 2\})$ whose entries are defined by $f_1(\mathbf{p}) = p_1 - 3(p_1 - {}^{1}\!/_{10})(p_1 - {}^{3}\!/_{5})(p_1 - {}^{9}\!/_{10})/2$ and $f_2(\mathbf{p}) = 1 - f_1(\mathbf{p})$ for all $\mathbf{p}$. Then f has fixed points at $\mathbf{p} = ({}^{1}\!/_{10}, {}^{9}\!/_{10})$, $\mathbf{p} = ({}^{3}\!/_{5}, {}^{2}\!/_{5})$, and $\mathbf{p} = ({}^{9}\!/_{10}, {}^{1}\!/_{10})$.

We focus on fixed points (or approximate fixed points) as a standard of honesty. To see why one may prefer reports that are fixed points, consider a case in which there are no strong guarantees (upper bounds) on $\|\boldsymbol{p} - f(\boldsymbol{p})\|$. Then the actual probability of an event, $f_i(\boldsymbol{p})$, could be much higher or lower than the reported probability p_i. This would prevent one from drawing any useful conclusions from the report. However, if $\boldsymbol{p} = f(\boldsymbol{p})$ or $\|\boldsymbol{p} - f(\boldsymbol{p})\|$ is small, then one can rely on the prediction $\boldsymbol{p}$ to guide decisions.

That being said, fixed points are not all one might care about, especially when it comes to potential superhuman oracle AIs. Ideally, we would want such systems to not think about how to influence the world at all [Armstrong and O'Rorke, 2017]. Alternatively, they should choose good fixed points over bad ones, hoping that such fixed points exist (we discuss preferences between different fixed points in Appendix B). Regardless, it is still important to understand whether and when fixed points are incentivized. For instance, if a model reports fixed points, one could try to use it only in situations in which a unique desirable fixed point exists.

Relation to performative prediction. As noted in the introduction, our setting is a special case of performative prediction [Perdomo et al., 2020]. In performative prediction, the goal is to find a model parameter that minimizes empirical risk for a classification or regression task, assuming that the choice of parameter can influence the data distribution. The loss-minimizing parameter when taking into account this influence is called performatively optimal. The analogue to fixed points in performative prediction are *performatively stable* predictions.

We indicate below when our results are analogous to results in the performative prediction setting. However, most of our results are unique to our setting. We take the perspective

[2] For example, in the case of three outcomes, the partial derivative at $(0.3, 0.4, 0.4)$ w.r.t. the first entry is the limit $\lim_{\epsilon \to 0}(G(0.3 + \epsilon, 0.4, 0.4) - G(0.3, 0.4, 0.4))/\epsilon$. But $G(0.3 + \epsilon, 0.4, 0.4)$ is not (necessarily) defined for positive (or negative) ϵ.

[3] Note that any other factor influencing the expert's belief $\boldsymbol{q}$ can be incorporated into f by marginalizing. For example, assume $\boldsymbol{q}$ is a function $\boldsymbol{q} = g(\boldsymbol{p}; X)$ where $\boldsymbol{p}$ is the expert's prediction and X is some environmental factor the expert is uncertain about. Then we can let $f(\boldsymbol{p}) := \mathbb{E}_X\left[g(\boldsymbol{p}, X)\right]$.

of a mechanism designer instead of taking a loss function as given. Moreover, we focus on fixed points instead of performative optima. In particular, we bound the quantity $\|\boldsymbol{p} - f(\boldsymbol{p})\|$ corresponding to the inaccuracy of predictions, which does not have a direct analogue in performative prediction. We give a more detailed comparison in Section 9.

Additional notation. We use $\mathbf{1}$ to denote the vector $(1, \ldots, 1)^\top \in \mathbb{R}^n$ and I to denote the identity matrix. We define $\operatorname{int}(\Delta(\mathcal{N})) := \{\boldsymbol{p} \in \Delta(\mathcal{N}) \mid \forall i\colon 0 < p_i < 1\}$ and use $\|\boldsymbol{x}\| := \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ to denote the Euclidean norm on $\mathbb{R}^n$.

4 INCENTIVES TO PREDICT NON-FIXED-POINTS

We begin by investigating whether an expert makes honest predictions, even in the presence of performativity. In performative prediction, it has been shown that performative optimality comes apart from performative stability (the analogous concept to a fixed point in our setting) [Perdomo et al., 2020, Izzo et al., 2021]. However, one may ask whether this is always the case or whether, e.g., some scoring function would prevent this.

We show that this is not the case: fixed points are in general not optimal. First, we show that for any strictly proper scoring rule there exist cases where a fixed point exists but the optimal prediction is not a fixed point. Afterwards, we show that when assuming differentiability and some reasonable distribution over f, optimal predictions are almost surely not fixed points.

Proposition 1. *Let S be any strictly proper scoring rule. For any interior fixed point $\boldsymbol{p}^* \in \operatorname{int}(\Delta(\mathcal{N}))$ there exists a function f with Lipschitz constant $L_f < 1$ and a unique fixed point at $\boldsymbol{p}^*$, such that there exists $\boldsymbol{p}' \neq \boldsymbol{p}^*$ with $S(\boldsymbol{p}', f(\boldsymbol{p}')) > S(\boldsymbol{p}^*, f(\boldsymbol{p}^*))$. That is, the unique fixed point of f is not performatively optimal.*

Note that since the function f has Lipschitz constant strictly smaller than 1, it represents a world that "dampens" the influence of the prediction, leading to a unique fixed point by Banach's fixed point theorem. It is interesting that the expert still prefers to make a prediction that is not a fixed point.

The above result raises the question whether a situation where fixed points are suboptimal is a niche counterexample or whether it is common. We show that under some relatively mild assumptions, the optimal prediction is almost surely not a fixed point. The intuition behind this result is that if a prediction $\boldsymbol{p}$ is an interior point and optimal, then $\nabla_{\boldsymbol{p}}(S(\boldsymbol{p}, f(\boldsymbol{p}))) = 0$. Using the Gneiting and Raftery characterization, we can show that this is a knife-edge case in which $g(\boldsymbol{p})^\top Df(\boldsymbol{p}) = 0$. Given sufficiently continuous distributions, this happens with probability 0. The conditions on the stochastic field $\{F(\boldsymbol{p})\}_{\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))}$ ensure this continuity, i.e., that the distributions over f as well as $Df(\boldsymbol{p})$ do not assign positive probability to any single point or subspace, hence almost never sampling the knife edge case. The condition would hold, e.g., for a Gaussian process with smooth kernel and mean functions (see Example 5 in Appendix A.3).

Theorem 2. *Let S be a twice differentiable strictly proper scoring rule. Let $\mathcal{F} := \{F(\boldsymbol{p})\}_{\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))}$ be a stochastic field with values in $\Delta(\mathcal{N})$ and let $Y(\boldsymbol{p}, \boldsymbol{v}) := (\Pi_{n-1} F(\boldsymbol{p}), \Pi_{n-1} \partial_{\boldsymbol{v}} F(\boldsymbol{p}))$ for $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ and $\boldsymbol{v} \in \mathcal{T} \cap S^{n-1}$. Assume that*

- *the sample paths $\boldsymbol{p} \rightsquigarrow F(\boldsymbol{p})$ are twice continuously differentiable*
- *for each $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ and $\boldsymbol{v} \in \mathcal{T} \cap S^{n-1}$, the random vector $Y(\boldsymbol{p}, \boldsymbol{v})$ has a joint density $h_{Y(\boldsymbol{p},\boldsymbol{v})}$ and there exists a constant C such that $h_{Y(\boldsymbol{p},\boldsymbol{v})} \leq C$ for all $\boldsymbol{p} \in \Delta(\mathcal{N}), \boldsymbol{v} \in S^{n-1} \cap \mathcal{T}$.*

Then, almost surely, there is no point $\boldsymbol{p} \in \operatorname{int}(\Delta(\mathcal{N}))$ such that $\boldsymbol{p} \in \arg\max_{\boldsymbol{p}'} S(\boldsymbol{p}', F(\boldsymbol{p}'))$ and $F(\boldsymbol{p}) = \boldsymbol{p}$.

5 BOUNDS ON THE DEVIATION FROM FIXED POINTS

In the previous section, we have shown that performatively optimal predictions are generally not fixed points, i.e., they inaccurately represent the expert's beliefs. But *how* inaccurate should we expect predictions to be, and what properties of S and f determine this inaccuracy? Assuming differentiability of f and S, this section provides upper bounds for the inaccuracy of optimal predictions $\boldsymbol{p}$ (i.e., $\|\boldsymbol{p} - f(\boldsymbol{p})\|$) and their distance from fixed points $\boldsymbol{p}^*$ (i.e., $\|\boldsymbol{p} - \boldsymbol{p}^*\|$). Note that, while the latter has a direct analogue in the performative prediction literature [Perdomo et al., 2020, Theorem 4.3], evaluating the inaccuracy of predictions only makes sense in our context where parameters are probability distributions.

For our bounds we will use the following notation. We use $\|A\|_{\mathrm{op}} = \max_{\boldsymbol{v} \in \mathcal{T}} \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|}$ for the operator norm of A on the tangent space. It is equal to A's largest singular value when seen as an automorphism on the tangent space. We use $A|_{\mathcal{T}} \succeq \gamma$ to denote that $\boldsymbol{v}^\top (A - \gamma I)\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathcal{T}$. If A is symmetric, this is equivalent to saying that the smallest eigenvalue of A on the tangent space is at least γ. Further, note that if g is a subderivative of G and $\|g(\boldsymbol{p})\| < L_G$ for all $\boldsymbol{p} \in \Delta(\mathcal{N})$, then L_G is a Lipschitz constant of G. Similarly, if $\|Df(\boldsymbol{p})\|_{\mathrm{op}} \leq L_f$ for all $\boldsymbol{p} \in \Delta(\mathcal{N})$, then L_f is a Lipschitz constant of f.

Theorem 3. *Let S be a strictly proper scoring rule, and let G, g as in the Gneiting and Raftery characterization (Theorem 1). Let $\boldsymbol{p} \in \Delta(\mathcal{N})$ and assume f, G, g are differentiable at $\boldsymbol{p}$. Assume $Dg(\boldsymbol{p})|_{\mathcal{T}} \succeq \gamma_{\boldsymbol{p}}$ for some $\gamma_{\boldsymbol{p}} > 0$.*

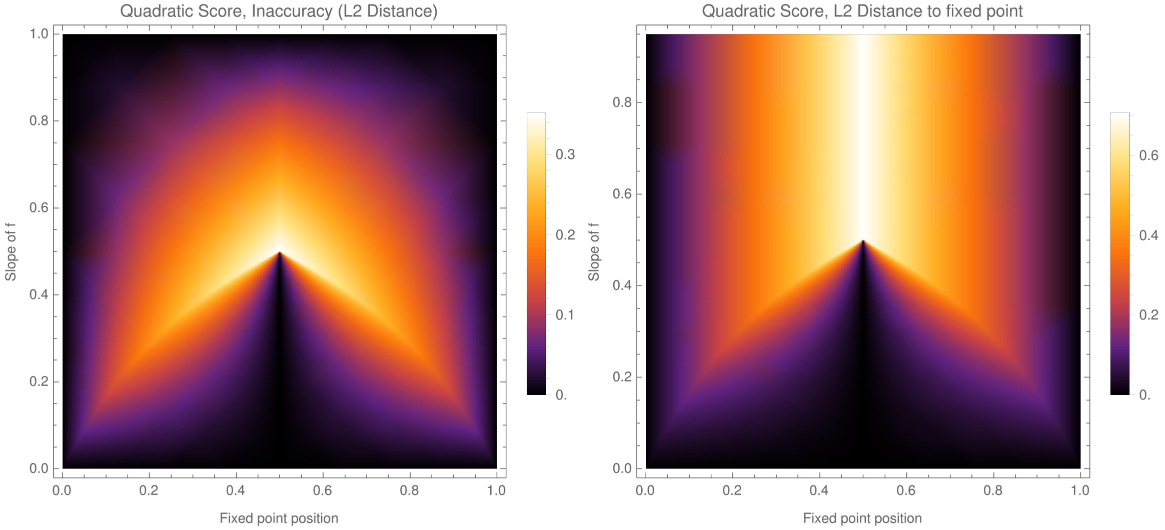

Figure 2: Heatmap of L2 distance of optimal prediction $\boldsymbol{p}$ to true probability distribution $f(\boldsymbol{p})$ (left) and to the fixed point $\boldsymbol{p}^*$ (right), depending on fixed point position p_1^* and α (slope of f), for the quadratic scoring rule.

Then whenever $\boldsymbol{p}$ is a performatively optimal report,

$$\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq \frac{\|Df(\boldsymbol{p})\|_{\text{op}}\|g(\boldsymbol{p})\|}{\gamma_{\boldsymbol{p}}}.$$

In particular, if f has Lipschitz constant L_f, G has Lipschitz constant L_G, and G is γ-strongly convex, then we have $\|\boldsymbol{p} - f(\boldsymbol{p})\| \leq \frac{L_f L_G}{\gamma}$.

In the case where f has Lipschitz constant $L_f < 1$, we can use the above results to derive a bound on how far the optimal report is from the (by Banach's fixed point theorem unique) fixed point.

Theorem 4. *Same assumptions as Theorem 3. Assume further that f has Lipschitz constant $L_f < 1$. Let $\boldsymbol{p}^*$ be the unique fixed point of f. Then for the performatively optimal report $\boldsymbol{p}$,*

$$\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq \frac{\|g(\boldsymbol{p})\|\|Df(\boldsymbol{p})\|_{\text{op}}}{(1 - L_f)\gamma_{\boldsymbol{p}}} \leq \frac{L_f L_G}{(1 - L_f)\gamma_{\boldsymbol{p}}}.$$

Note that the assumption that $L_f < 1$ ensures that f's fixed point is unique by Banach's fixed point theorem. Without $L_f < 1$, no trivial bound holds, as we show in Proposition 3 in Appendix A.6.

This bound is analogous to a bound in [Perdomo et al., 2020, Theorem 4.3]. Our bound differs in that we use Euclidean distance instead of Wasserstein distance to measure the sensitivity of f to the choice of report. Moreover, assuming a L_ℓ-Lipschitz and γ-strictly convex loss function ℓ, their bound depends on the ratio $\frac{L_\ell}{\gamma}$. We instead bound this distance against the ratio $\frac{\|g(\boldsymbol{p})\|}{\gamma_{\boldsymbol{p}}}$, which will allow us to minimize the bound in the two-outcome case by using exponential functions (Theorem 5). This would not be possible when assuming γ-strict convexity, since there exist no functions that globally make the ratio $\frac{L_\ell}{\gamma}$ arbitrarily small. Perdomo et al. [2020] show that their bound can be made small by regularizing the loss function, but this would be undesirable in our setting, since regularized scoring rules would be improper and thus cease to incentivize honest reports even for constant f.

Example 4 (Bound for the quadratic scoring rule)**.** Consider the quadratic scoring rule $S(\boldsymbol{p}, i) = 2p_i - \|\boldsymbol{p}\|_2$. Note that we can represent this in Gneiting and Raftery's characterization with $G(\boldsymbol{p}) = \|\boldsymbol{p}\|^2$ and $g(\boldsymbol{p}) = 2\boldsymbol{p} - \frac{2}{n}\mathbf{1}$. Thus, $Dg(\boldsymbol{p}) = 2I$, where I is the identity matrix. Hence $Dg(\boldsymbol{p}) \succ 2$. Further, $\|g(\boldsymbol{p})\|_2 = 2\|\boldsymbol{p} - \frac{1}{n}\mathbf{1}\|$. Thus, for f with Lipschitz constant L_f, Theorem 3 implies that for the optimal report $\boldsymbol{p}$ we have that $\|f(\boldsymbol{p}) - \boldsymbol{p}\| \leq L_f\|\boldsymbol{p} - \frac{1}{n}\mathbf{1}\| \leq L_f\sqrt{(n-1)/n}$. If $L_f < 1$, then by Theorem 4 we further have $\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq \frac{L_f}{1-L_f}\|\boldsymbol{p} - \frac{1}{n}\mathbf{1}\| \leq \frac{L_f}{1-L_f}\sqrt{(n-1)/n}$.

6 APPROXIMATE FIXED-POINT PREDICTION WITH THE RIGHT PROPER SCORING RULES?

The above results show that depending on the scoring rule we can obtain bounds on the accuracy of performatively optimal predictions. Can we make these bounds arbitrarily small by choosing an appropriate scoring rule, e.g., one that makes $\|g(\boldsymbol{p})\|/\gamma_{\boldsymbol{p}}$ very small at each point? In this section, we show that the answer is yes in the two-outcome case and no in the general case.

Theorem 5. *Consider the case of two outcomes, i.e., let $\mathcal{N} = \{1, 2\}$. Let $L_f \in \mathbb{R}$ and $\epsilon > 0$. Then there exists a* *scoring rule S s.t. under any f with Lipschitz constant L_f, any optimal report $\mathbf{p}$ satisfies $\|\mathbf{p} - f(\mathbf{p})\| \leq \epsilon$. If $L_f < 1$, then there also exists a scoring rule that additionally ensures that under any f with Lipschitz constant L_f, any optimal report satisfies $\|\boldsymbol{p} - \boldsymbol{p}^*\| \leq \epsilon$, where $\boldsymbol{p}^*$ is the (unique) fixed point of f.*

Note that if there are multiple fixed points, then S still induces preferences between—approximately—predicting these fixed points. In particular, because $S(\boldsymbol{p}, \boldsymbol{p})$ is convex, the performatively optimal fixed point will either be the one that maximizes or the one that minimizes p_1 among the fixed points. This may be undesirable as the expert still has a strong incentive other than (though compatible with) honest prediction. We discuss this in more detail in Appendix B.

Can arbitrarily good bounds be achieved with *practical* proper scoring rules? Our proof of Theorem 5 uses exponential scoring rules with $g(\boldsymbol{p}) = (e^{L_f p_1/(\sqrt{2}\epsilon)}, -e^{L_f p_1/(\sqrt{2}\epsilon})^\top$. For high K, this scoring rule seems impractical, because the stakes vary greatly over the interval. For example, $S((2/3 + \epsilon, 1/3 - \epsilon), (2/3, 1/3))/S((1/2 + \epsilon, 1/2 - \epsilon), (1/2, 1/2)) = e^{L_f/(6\sqrt{2}\epsilon)}$. Hence, as we increase L_f/ϵ, it becomes exponentially more important for the expert to predict accurately near $2/3$ than to predict accurately near $1/2$. In particular, an AI model trained with this scoring rule may be much worse at predicting probabilities near $1/2$ than near $2/3$. Similarly, it is unrealistic to reward a human expert with, say, millions of dollars near $2/3$ and with just a few cents near $1/2$. Unfortunately, it turns out that all possible scoring rules that achieve bound ϵ under Lipschitz constant L_f have this undesirable property, though the exact bound turns out somewhat complicated.

Theorem 6. *Suppose S is a proper scoring rule s.t. for some $\epsilon, L_f > 0$ we have that whenever f is L_f-Lipschitz, the optimal report $\mathbf{p}$ satisfies $\|f(\boldsymbol{p}) - \boldsymbol{p}\| < \epsilon$. Let $3\epsilon \leq p_l \leq p_h \leq 1 - 4\epsilon$ and $\delta = \epsilon/(L_f + 1)$. Then the ratio of the supremum and infimum over $p_1 \in [p_l, p_h]$ of $S((p_1 + 4\delta, 1 - p_1 - 4\delta), (p_1, 1 - p_1)) - S((p_1, 1 - p_1), (p_1, 1 - p_1))$ is at least*

$$\frac{L_f}{2L_f + 6}\left(3\frac{L_f + 1}{L_f + 3}\right)^{(L_f+1)(p_h - p_l)/(8\epsilon) - 5/2}.$$

In particular, for fixed positive L_f, this term is exponential in $1/\epsilon$ and for fixed positive ϵ it is exponential in L_f.

Intuitively, the assumption on S is that it ensures small accuracy bounds of ϵ for functions with Lipschitz constant L_f. Now note that $|S((p_1 + 4\delta, 1 - p_1 - 4\delta), (p_1, 1 - p_1)) - S((p_1, 1 - p_1), (p_1, 1 - p_1))|$ is the cost to the expert of misreporting by 4δ when the true distribution is $(p_1, 1 - p_1)$. If this term is large, then the expert cares a lot about not misreporting by 4δ, and if the term is small, the expert does not mind misreporting much. Our result shows that the value of this term is much larger for some p_1 than it is for others, i.e., that for some probabilities p_1 the expert cares a lot more about accurately reporting p_1 than it does for other values of p_1. Our theorem puts a lower bound on the ratio between the lowest and largest possible values of that term. In particular, this does not hinge on probabilities p_1 near 0 or 1 and holds even if we restrict attention to probabilities between, say, $1/4$ and $3/4$.

Theorem 5 shows that in the binary prediction case, given a Lipschitz constant L_f for the environment, we can achieve arbitrarily good bounds ϵ on the inaccuracy of the performatively optimal report. Unfortunately, this ceases to be possible in the many-outcome case. In that case, if all we know about f is that it has Lipschitz constant L_f, there is some error ϵ, linear in L_f as $L_f \to 0$, that we must allow regardless of what strictly proper scoring rule we use.

Theorem 7. *For any Lipschitz constant L_f, for $\epsilon > 0$ sufficiently small, there is no proper scoring rule S for the three-outcome case that achieves the following property: Whenever f is L_f-Lipschitz, there is some performatively optimal report $\mathbf{p}$ with $\|f(\boldsymbol{p}) - \mathbf{p}\| \leq \epsilon$. In particular, there exists some function $\epsilon(L_f)$ with $\epsilon(L_f) \sim cL_f$ as $L_f \to 0$ for some fixed constant c, s.t. the above property cannot be achieved with $\epsilon = \epsilon(L_f)$. Thus, the best achievable bound is in $\Omega(L_f)$ as $L_f \to 0$, i.e. scales at least linearly with L_f in the limit.*

7 NUMERICAL SIMULATIONS

In this section, we provide some numerical simulations for the Brier score, to see how inaccurate performatively optimal predictions might be in practice. Throughout, we consider only affine-linear functions f. This in particular means that all functions f have operator norms between 0 and 1 and aside from degenerate cases a unique fixed point. The Mathematica notebook for our experiments (including some interactive widgets) is available at `https://github.com/johannestreutlein/scoring-rules-performative`. Although our experiments are set in toy models with linear f and small sets of outcomes, they provide an initial estimate of the degree to which predictions can be off, depending on how much influence the expert can exert using their prediction.

7.1 BINARY PREDICTION

Experimental setup. We begin with the binary prediction case, i.e., $\mathcal{N} = \{1, 2\}$. We consider f to be affine linear with slope α and fixed point $\boldsymbol{p}^* \in \Delta(\mathcal{N})$, thus yielding the functional form $f(\boldsymbol{p}) := \boldsymbol{p}^* + \alpha(\boldsymbol{p} - \boldsymbol{p}^*)$ for all $\boldsymbol{p} \in \Delta(\{1, 2\})$. Note that for all $\alpha \in [0, 1]$ and all $\boldsymbol{p}^* \in \Delta(\mathcal{N})$, a function thus defined is indeed a function $\Delta(\mathcal{N}) \to \Delta(\mathcal{N})$. For

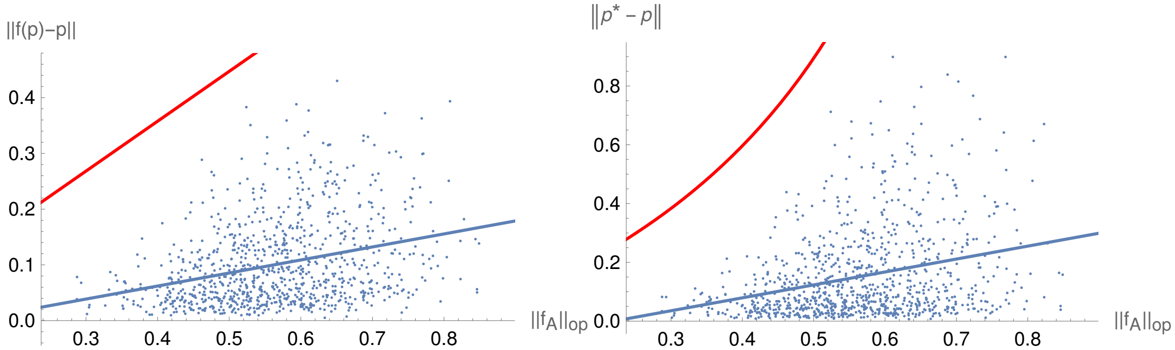

Figure 3: Scatter plots showing the L2 inaccuracy (left) and the distance to a fixed point (right) of the performatively optimal reports against the operator norm of A in our experiments. In both plots, each point corresponds to a run of the experiments. The blue lines are found by linear regression on the points. The red lines are the bound given in Example 4 as a function of the Lipschitz constant L_f.

$\alpha < 0$, whether f is a function $\Delta(\mathcal{N}) \to \Delta(\mathcal{N})$ depends on $\boldsymbol{p}^*$. We restrict attention to $\alpha \in [0, 1]$ for simplicity.

Graphing inaccuracy and distance to fixed points. In Figure 2 (left), we plot the inaccuracy $\|\boldsymbol{p} - f(\boldsymbol{p})\|$ of the performatively optimal report $\boldsymbol{p}$ against α, p_1^*. In Figure 2 (right), we plot the L2 distance $\|\boldsymbol{p}^* - \boldsymbol{p}\|$ of the performatively optimal report $\boldsymbol{p}$ to the fixed point $\boldsymbol{p}^*$. For that plot we limit α to the range $[0, 0.95]$, because of instability at $\alpha \approx 1$. Note that relatively high inaccuracies can be found at various qualitatively different points in the graphs, even when the slope of f is small, i.e., when the oracle has little influence on the environment.

Assessing our bounds. To evaluate our bounds, we maximize distances across possible choices of fixed points $\boldsymbol{p}^* \in \Delta(\{1, 2\})$, and plot the maximal inaccuracy of the optimal prediction as well as the maximal distance from a fixed point in Figure 1. We compare to both theoretical bounds from Example 4, i.e., $\|p - f(p)\|_2 \leq \alpha/\sqrt{2}$ and $\|p - p^*\|_2 \leq \alpha/((1-\alpha)\sqrt{2})$.

For both quadratic and log scoring rule (results in Appendix C), our theoretical bounds are tight for slopes $\alpha \leq 0.5$. For higher slopes, inaccuracy goes down, as the function $f(p)$ becomes closer to the identity function, and optimal predictions are bounded in $[0, 1]$.

7.2 HIGHER-DIMENSIONAL PREDICTION

Experimental setup. Next, we turn to higher-dimensional predictions. We consider a model with five possible outcomes and linear $f \colon p \mapsto Ap$ for $A \in \mathbb{R}^{n \times n}$. f is an automorphism on the simplex if and only if all of its columns are in the simplex. We hence randomly generate the matrix A by sampling each column uniformly from the simplex. Note that A is the Jacobian of f at every point.

For each f_A thus created, we first find the performatively optimal report $\boldsymbol{p}$ and the fixed point $\boldsymbol{p}^*$. We then record the following quantities: the operator norm of f_A; the distance of the fixed point distribution to the uniform distribution $\|\boldsymbol{p}^* - \frac{1}{n}\mathbf{1}\|$; the distance of the optimal report to the uniform distribution $\|\boldsymbol{p} - \frac{1}{n}\mathbf{1}\|$; the distance of the performatively optimal report to the fixed point $\|\boldsymbol{p}^* - \boldsymbol{p}\|$; the inaccuracy of the performatively optimal report $\|f(\boldsymbol{p}) - \boldsymbol{p}\|$. We are interested in how the second two items depend on the first two. We are also interested in how tight our bounds (from Example 4) are.

We collected 1000 random functions f_A, but aborted 52 runs because they didn't terminate within 120 seconds, leaving us with 948 data points.

Inaccuracy. Figure 3 (left) plots the L2 inaccuracy (i.e., the distances $\|\boldsymbol{p} - f(\boldsymbol{p})\|$). The blue line shows the best linear fit to the data points, which is given by $-0.0314 + 0.234x$, whereas our bound is $2L_f/\sqrt{5} \approx 0.8944L_f$. The average L2 inaccuracy is 0.100 with a standard deviation of 0.0770. The quartiles are $0.0419, 0.0759, 0.138$. The correlation between the operator norm of f_A and $\|\boldsymbol{p} - f(\boldsymbol{p})\|$ is 0.312.

Distance to fixed points. Figure 3 (right) plots the L2 distance to the fixed point against the operator norm of f_A. The linear best fit (blue line) is given by $-0.0966 + 0.440x$, whereas our bound is $2L_f/(\sqrt{2}(1 - L_f)) \approx 0.8944L_f/(1 - L_f)$. The average L2 distance to the fixed point is 0.152 with a standard deviation of 0.154 and quartiles $0.0442, 0.0915, 0.210$. The correlation between the operator norm of f_A and $\|\boldsymbol{p}^* - \boldsymbol{p}\|$ is 0.294.

The role of the location of the fixed point. The graphs for the binary prediction case show that the location of the fixed point matters a lot for the accuracy of optimal reports (though the direction of the effect depends on the slope of f). A similar effect can be observed in the many outcome case. In fact, the effect of the location of the fixed point is actually stronger (though less reliable) than the effect of the operator norm of f_A. We provide more detail in Appendix C.2.2.

Loose bounds, tight bounds. Figure 3 show that (in contrast to the binary prediction case), our bounds in terms of the operator norm of f are typically quite loose. For example, the average slack of the inaccuracy bound is 0.404 with a standard deviation of 0.0998 and quartiles $0.337, 0.400, 0.471$. Recall from Example 4 that in addition to bounds in terms of L_f alone we have bounds in terms of L_f and $\|\boldsymbol{p} - \frac{1}{n}\mathbf{1}\|$. These bounds are much tighter with an average slack of 0.0644 with a standard deviation of 0.0597 and quartiles $0.0274, 0.0487, 0.0830$.

Discussion. Based on our simulations, misprediction in the five outcome case seems similarly problematic as in the binary prediction case. In contrast to the binary case, the bounds in terms of L_f are quite loose. The bounds in terms of $\|\boldsymbol{p} - \frac{1}{n}\mathbf{1}\|$ are much tighter. Note that because these bounds depend on the performatively optimal report, they can only be derived a posteriori once a report has been submitted. As in the two-outcome case, both the location of the fixed point and the operator norm/slope of f matter a lot for accuracy and distance to fixed point of the performatively optimal report.

8 FIXED POINTS VIA ALTERNATIVE NOTIONS OF OPTIMALITY

Here, we focus on alternative settings that lead to accurate predictions and do not induce preferences over fixed points. The idea behind all of them is that, instead of optimizing $\boldsymbol{p}$ and $f(\boldsymbol{p})$ jointly, we keep $\boldsymbol{q} := f(\boldsymbol{p})$ fixed while choosing a prediction $\boldsymbol{p}$ to maximize $S(\boldsymbol{p}, \boldsymbol{q})$. Repeating this procedure leads to honest predictions, where the choice of fixed point depends on contingent facts such as initialization, instead of being chosen to maximize $S(\boldsymbol{p}, \boldsymbol{p})$. An AI model using this procedure could be safer, because its predictions are honest, and because it does not optimize its choice of fixed point for any goal. In this section we give a summary of a more detailed treatment with formal results in Appendix D.

Performative stability. Alternatives to performative optimality have been discussed in the performative prediction literature. Translated into our setting, a prediction $\boldsymbol{p}^*$ is called *performatively stable* if $\boldsymbol{p}^* \in \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, f(\boldsymbol{p}^*))$. This implies $\boldsymbol{p}^* = f(\boldsymbol{p}^*)$ whenever S is strictly proper, so performative stability is equivalent to being a fixed point.

Repeated risk minimization and gradient descent. Perdomo et al. [2020] consider learning algorithms that converge to performatively stable points, including repeated risk minimization and repeated gradient descent. In repeated risk minimization, we repeatedly update predictions via $\boldsymbol{p}_{t+1} := \arg\max_{\boldsymbol{p}} S(\boldsymbol{p}, f(\boldsymbol{p}_t))$. Repeated gradient descent instead updates predictions via gradient descent on this objective. There also exist stochastic gradient descent versions of these algorithms [Mendler-Dünner et al., 2020]. All of these schemes lead to stable points under appropriate conditions. We include a convergence proof for repeated gradient descent in our setting in Appendix D.2.

No-regret learning and prediction markets. We also provide results for no-regret learning (Appendix D.4) and prediction markets (Appendix D.5). We introduce a no-regret learning setting and show that policies have sublinear regret if and only if they have sublinear prediction error. This differs from the setting considered by Jagadeesan et al. [2022], in which no-regret policies converge to performatively optimal predictions. Next, we provide a prediction market model and show that, if the weight of each trader in the market is small, equilibrium predictions by the market are close to fixed points. This is analogous to a result by Hardt et al. [2022] bounding the distance of a market equilibrium from performatively stable points.

9 RELATED WORK

Performative prediction. In performative prediction, the goal is to find a model parameter $\theta \in \mathbb{R}^d$ that minimizes an expected loss $\mathbb{E}[\ell(Z; \theta)]$ where Z is a stochastic sample, usually a pair of input and target, $Z = (X, Y)$. Unlike in the vanilla supervised learning setting, $Z \sim \mathcal{D}(\theta)$ is sampled from a distribution $\mathcal{D}(\theta)$ that itself depends on the chosen model parameter. Performatively optimal parameters are defined via $\theta_{\text{PO}} \in \arg\min_\theta \mathbb{E}_{Z \sim \mathcal{D}(\theta)}[\ell(Z; \theta)]$, and the definition of performatively stable parameters is $\theta_{\text{PS}} \in \arg\min_\theta \mathbb{E}_{Z \sim \mathcal{D}(\theta_{\text{PS}})} \ell(Z; \theta)$. In general, performatively stable and optimal parameters can differ [Perdomo et al., 2020, Ex. 2.2].

Our setting could be seen as a special case in which θ is a single distribution $\boldsymbol{p}$, data points are discrete outcomes y, and the distribution $\mathcal{D}(\theta)$ is given by $f(\boldsymbol{p})$. Unlike in the general performative prediction setting, we can determine the accuracy of a prediction $\boldsymbol{p}$ as the distance from the distribution $f(\boldsymbol{p})$ (see Theorem 3), we can characterize predictions as honest if they are fixed points, and loss functions can be characterized as proper if they incentivize honest reports. As mentioned in Section 8, performatively stable points are fixed points and are thus a more desirable solution concept in our setting. There are some performative prediction settings in which performative optima can also be seen as manipulative and undesirable, such as in recommendation algorithms [Hardt et al., 2022]. However, as far as we are aware, we are the first to link performative stability to honesty in prediction.

Scoring rules. While the literature on scoring rules generally assumes that predictions are not performative, a few authors in this literature have studied agents manipulating the world *after* making a prediction Shi et al. [2009], Oka et al. [2014]. To our knowledge, the cases discussed do not involve agents influencing the world directly through their predictions. Chan [2022] introduce performative probabilistic predictions using scoring rules. However, they focus on particular functional forms of f and binary predictions and do not provide a more general analysis. Another related setting in which it has been shown that no proper scoring rules exist is that of second-order prediction, in which experts report distributions over first-order distributions to express epistemic uncertainty [Bengs et al., 2023].

AI oracles. Issues with performativity have been mentioned in the literature on AI predictors or oracles Armstrong and O'Rorke [2017]. Most prior work has focused on alleviating performativity altogether, e.g., by making the oracle predict *counterfactual worlds* it cannot influence. We are not aware of any prior work on specifically the question of whether AI oracles would be incentivized to output fixed points at all.

Decision scoring rules and decision markets. The literature on decision scoring rules and decision markets considers a setting in which experts make predictions about what would happen if a decision maker were to pursue one course of action or another. The decision maker then chooses based on these predictions, making the predictions performative. As shown by Othman and Sandholm [2010], the expert may thus be incentivized to mispredict when subject to a proper scoring rule. However, this literature typically takes the perspective of the decision maker and thus assumes some knowledge of f. For example, Othman and Sandholm [2010] and Oesterheld and Conitzer [2020b] show that the scoring rule S must be chosen to align in some sense with the decision maker's utility function (and thus f). Chen et al. [2014] propose that the decision maker could randomize to set good incentives, which in our setting would entail manipulating f.

Epistemic decision theory. A related topic in philosophy is *epistemic decision theory*. In particular, Greaves [2013] introduces several cases in which outcomes depend on the agent's credences and compares the verdicts of different epistemic decision theories (such as an evidential and a causal version). While some of Greaves' examples involve agents knowably adopting incorrect beliefs, they require joint beliefs over several propositions, and Greaves only considers individual examples. We instead consider only a single binary prediction and prove results for arbitrary scoring rules and relationships between predictions and beliefs.

Honest and truthful AI. Another related topic is honest and truthful AI [Evans et al., 2021]. In our setting, an AI that reports an inaccurate prediction to achieve a higher score would be dishonest. Evans et al. [2021] discuss issues around training AIs to be truthful and honest, such as difficulties in judging truth. However, they do not explore performativity or proper scoring rules. We simplify our analysis by assuming that a ground truth exists and can be judged objectively. Burns et al. [2022] discuss extracting latent knowledge from AIs without relying on incentivizing honest reporting, but also do not address performativity.

10 CONCLUSION AND FUTURE WORK

If predictions cannot influence which outcome occurs, then strictly proper scoring rules incentivize experts (humans or AI systems) to report honest predictions. This fails if predictions are performative. We showed that, in general, strictly proper scoring rules do not incentivize accurate predictions in a performative prediction setting. We analyzed this inaccuracy quantitatively and gave upper bounds on inaccuracy. We showed that in the case of binary prediction, there exist scoring rules that incentivize arbitrarily accurate predictions. In contrast, for more than two outcomes, it is not possible to achieve arbitrarily strong bounds on accuracy. Our numerical simulations in a toy setting confirm that our bounds are tight in some situations and that inaccurate performative predictions are common. Finally, we showed that by using other types of objectives, such as minimizing regret, we can build AI models that predict fixed points.

We hope that future work will shed further light on practical and safe uses of AI systems as predictors, i.e., oracle AIs. First, some of our bounds could probably be improved or generalized (to non-differentiable f, G). Second, it would be valuable to have more specific models of f. Precise models of f may allow for stronger results [cf. Othman and Sandholm, 2010, Oesterheld and Conitzer, 2020b]. Third, we take a simplistic view of safety: we take it that incentives to predict honestly are good and that other incentives are problematic. We hope that future work will augment our analysis with more fine-grained models of safety. For example, a common safety concern is power-seeking behavior [Omohundro, 2008, Turner et al., 2021]. One could similarly ask to what extent performative oracle AI will spend compute to improve its ability to influence the world (cf. discussions of information acquisition, e.g. Osband, 1989; Neyman et al., 2021; Li et al., 2022; Oesterheld and Conitzer, 2020a). Lastly, we are interested in theoretical and experimental evaluations of the practicality of different safe oracle AI designs and training setups.

Acknowledgements

CO acknowledges funding from the Cooperative AI Foundation, Polaris Ventures (formerly Center for Emerging Risk Research) and Jaan Tallinn's donor-advised fund at Founders Pledge. JT and RH carried out most of this work as part of the SERI MATS program under the mentorship of Evan Hubinger (JT, RH) and Leo Gao (RH). JT is grateful for support by an Open Phil AI Fellowship and an FLI PhD Fellowship. We sincerely thank four anonymous reviewers whose insightful comments helped us improve our paper. We are also indebted to Meena Jagadeesan, Erik Jenner, Adam Jermyn, and Marius Hobbhahn for their valuable discussions and feedback, and to Alexander Pan and Bastian Stern for pointing us to the relevant related literature.

References

S. Armstrong. Risks and mitigation strategies for oracle ai. In *Philosophy and Theory of Artificial Intelligence*, pages 335–347. Springer, 2013.

S. Armstrong and X. O'Rorke. Good and safe uses of AI oracles. *arXiv preprint arXiv:1711.05541*, 2017.

S. Armstrong, A. Sandberg, and N. Bostrom. Thinking inside the box: Controlling and using an oracle ai. *Minds and Machines*, 22(4):299–324, 2012.

V. Bengs, E. Hüllermeier, and W. Waegeman. On second-order scoring rules for epistemic uncertainty quantification. *arXiv preprint arXiv:2301.12736*, 2023.

N. Bostrom. *Superintelligence*. Oxford University Press, 2014.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1 1950.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

A. Carvalho. An overview of applications of proper scoring rules. *Decision Analysis*, 13(4):223–242, 2016.

A. Chan. Scoring rules for performative binary prediction. *arXiv preprint arXiv:2207.02847*, 2022.

Y. Chen, I. A. Kash, M. Ruberry, and V. Shnayder. Eliciting predictions and recommendations for decision making. In *ACM Transactions on Economics and Computation*, volume 2, chapter 6. 6 2014.

O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14:107–114, 1952.

H. Greaves. Epistemic decision theory. *Mind*, 122(488): 915–952, 2013.

M. Hardt, M. Jagadeesan, and C. Mendler-Dünner. Performative power. In *NeurIPS*, 2022.

Z. Izzo, L. Ying, and J. Zou. How to learn when data reacts to your model: performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4641–4650. PMLR, 2021.

M. Jagadeesan, T. Zrnic, and C. Mendler-Dünner. Regret minimization with performative feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9760–9785. PMLR, 2022.

Y. Li, J. D. Hartline, L. Shan, and Y. Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 988–989, 2022.

J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences of the United States of America*, 42:654–655, 9 1956.

C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. *NeurIPS*, 33:4929–4939, 2020.

E. Neyman, G. Noarov, and S. M. Weinberg. Binary scoring rules that incentivize precision. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 718–733, 2021.

R. Ngo, L. Chan, and S. Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

C. Oesterheld and V. Conitzer. Minimum-regret contracts for principal-expert problems. In *Proceedings of the 16th Conference on Web and Internet Economics (WINE)*. 2020a.

C. Oesterheld and V. Conitzer. Decision scoring rules. In *International Workshop on Internet and Network Economics*, page 468, 2020b.

M. Oka, T. Todo, Y. Sakurai, and M. Yokoo. Predicting own action: Self-fulfilling prophecy induced by proper scoring rules. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

S. M. Omohundro. The basic ai drives. In *Proceedings of the 2008 conference on Artificial General Intelligence: Proceedings of the First AGI Conference*, pages 483–492. IOS Press, 2008.

K. Osband. Optimal forecasting incentives. *Journal of Political Economy*, 97(5):1091–1112, 10 1989.

A. Othman and T. Sandholm. Decision rules and decision markets. In *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010), van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada*, pages 625–632. 2010.

J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7599–7609. PMLR, 2020.

S. J. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66:783–801, 12 1971.

P. Shi, V. Conitzer, and M. Guo. Prediction mechanisms that do not incentivize undesirable actions. In *International Workshop on Internet and Network Economics*, pages 89–100. Springer, 2009.

A. Turner, L. Smith, R. Shah, A. Critch, and P. Tadepalli. Optimal policies tend to seek power. *NeurIPS*, 34:23063–23074, 2021.