Joint Scoring Rules: Competition Between Agents Avoids Performative Prediction

Rubi Hudson

[1]University of Toronto

Abstract

In a decision-making scenario, a principal could use conditional predictions from an expert agent to inform their choice. However, this approach would introduce a fundamental conflict of interest. An agent optimizing for predictive accuracy is incentivized to manipulate their principal towards more predictable actions, which prevents that principal from being able to deterministically select their true preference. We demonstrate that this impossibility result can be overcome through the joint evaluation of multiple agents. When agents are made to engage in zero-sum competition, their incentive to influence the action taken is eliminated, and the principal can identify and take the action they most prefer. We further prove that this zero-sum setup is unique, efficiently implementable, and applicable under stochastic choice. Experiments in a toy environment demonstrate that training on a zero-sum objective significantly enhances both predictive accuracy and principal utility, and can eliminate previously learned manipulative behavior.

Technical Appendx —
https://arxiv.org/src/2412.20732/anc

Code —
https://github.com/rubi-hudson/Joint-Scoring-Rules

Introduction

Large Language Models (LLMs) excel at pattern recognition and predicting text. As capabilities improve, this predictive ability will be applied more frequently to real-world outcomes. This could include predicting human preferences, as is used in reinforcement learning from human feedback (RLHF) (Christiano et al. 2017), or predicting outcomes for multiple actions when only one action can be taken. Restricting highly capable AI systems to only make predictions, rather than autonomously pursue goals, is one proposal for AI alignment known as Oracle AI (Armstrong, Sandberg, and Bostrom 2012; Armstrong 2013).

A potential risk of using AI to make predictions is that the very act of making a prediction can affect the outcome, in a phenomenon known as performative prediction (Perdomo et al. 2020). For example, predictions of inflation affect consumer behavior and thus inflation, predictions of crime affect police deployments and thus crime, and predictions of mortality affect healthcare usage and thus mortality.

When non-performative predictions are evaluated using a strictly proper scoring rule, the only optimal response is honesty (Gneiting and Raftery 2007), where honesty means an agent reports their actual beliefs. However, all such rules give higher expected scores when the underlying distribution of outcomes is more extreme. In the performative case, such rules then incentivize making the distribution more extreme. (Oesterheld et al. 2023).

This introduces two risks. The first is that such manipulation could push the world towards becoming more predictable, with potentially negative effects on inflation, crime, healthcare, or other outcomes. In the worst case, it could pose an existential risk — a dead human is more predictable than a live one. (Hubinger et al. 2023). The second risk is that optimal performative predictions are typically not stable upon reflection (Oesterheld et al. 2023), which dramatically lowers their usefulness.

Still, one might ask how performative prediction could arise if we do not intentionally run a gradient running through a prediction's impact on outcomes? We identify four main pathways for this to occur:

1. If a model implements search, either by design or as learned behavior (Hubinger et al. 2019), and chooses their prediction based on expected score.
2. If there are selection pressures other than gradient descent, such as population based training (Krueger, Maharaj, and Leike 2020) or economic competition (Hendrycks 2023).
3. If there is generalization (Shah et al. 2022), from only historical training data to an environment where the model can influence the outcome.
4. If powerful models are developed through a process other than gradient-based updating.

One method for avoiding performativity is to elicit predictions conditional on various actions that can be taken in response to the prediction. As this reaction is the causal pathway by which a prediction affects outcomes, conditioning on it removes the performative aspect.

Unfortunately, this approach introduces a new type of performativity. The predictor can influence which action gets taken by dishonestly reporting certain predictions. In fact, it is impossible for a decision maker to always take their most

preferred action when eliciting conditional predictions from an agent using a symmetric scoring rule (Othman and Sandholm 2010). The symmetry of the scoring rule, meaning it is invariant to the specifics of outcomes, is important because it avoids the need to establish preferences over all outcomes. This is a famously challenging aspect of the AI alignment problem, due to the difficulty of specifying one's full values (Hadfield-Menell and Hadfield 2019).

As a motivating example, consider an ICU at a hospital that assigns extra care to patients with a predicted chance of mortality over 70%. For a type of patient with a 90% baseline chance of mortality that drops to a 20% chance if they receive extra care, using gradient descent to train the predictor would not converge, instead fluctuating across the 70% threshold (Bell et al. 2021). Importantly though, if the predictor is somehow trained to make optimal prediction, this will be exactly 70%, which causes the highest mortality. When conditional predictions are used instead, the optimal pair is 90% without extra treatment and >90% with, so no treatment looks better. Taking that action and evaluating that prediction results in the highest score for the predictor. Only by using our proposed method can we make honest predictions optimal, allowing the ICU to minimize mortality.

Our Contribution

Our contribution is to show that when jointly evaluating two or more predictors, it is possible to elicit honest predictions using a symmetric scoring rule and to use this info to deterministically take the best action. This enables the safe use of conditional predictions from powerful AI systems, either for their own sake or to avoid performative prediction.

Our primary result is a theorem showing that when predictors engage in a zero-sum competition for accuracy, they have no incentive to influence which action gets taken, and a decision maker can exploit this to elicit honest conditional predictions. This is paired with a uniqueness theorem, showing that only zero-sum competition can make this possible.

Supplementary results address obstacles to implementing the mechanism in practice. We show large spaces of possible actions can be efficiently searched to identify the optimal choice. Additionally, we extend our main results to stochastic choice from the decision maker, which can add further incentives for honesty regarding untaken actions.

Finally, we use experiments in a toy environment to show that models trained with a zero-sum objective do not become performative, even in environments incentivizing it. We also demonstrate that zero-sum training untrains performativity from a model faster and to a larger degree than directly removing the incentives for performativity.

Related Work

The literature on eliciting honest predictions using proper scoring rules is well established (Brier 1950; Good 1952; Savage 1971; Gneiting and Raftery 2007). However, it largely depends on the assumption that making a prediction does not impact the outcome.

Concern about predictions affecting outcomes with respect to Oracle AI began in Armstrong and O'Rorke (2017), with a more rigorous formulation and the term "performative prediction" being introduced in Perdomo et al. (2020). This was followed up with several papers concerning performativity in a single prediction, and largely focused on stability, optimal predictive accuracy, and training methods (Mendler-Dünner et al. 2020, Izzo, Ying, and Zou 2021; Hardt, Jagadeesan, and Mendler-Dünner 2022). We focus on performativity across multiple conditional predictions, and address the impact of performative predictions on the welfare of those who elicit them.

Using conditional predictions to choose an action was introduced as a decision problem in Othman and Sandholm (2010), which went on to prove the impossibility result for using them to take optimal actions. This is the main paper our work responds to, by showing that these results can be overcome using multiple predictors.

Chen et al. (2011) shows that accurate conditional predictions can be elicited if the decision maker is willing to randomize with full support over all actions. However, a human decision maker often cannot commit to randomly taking sufficiently bad actions, and they may not know the exact probabilities they assign to each action as required. In an ML context, arbitrarily small probabilities require arbitrarily large amounts of training data.

Oesterheld and Conitzer (2020) shows that a decision maker can take the best action by having the expert choose the action, and reward them proportionally to the resulting utility. This requires specifying their utility function, which is often difficult in general environments. This approach can also fail when outcomes modify the decision maker's utility function or ability to report it. Additionally, the information gained from the predicted probabilities can have its own value.

Background and Definitions

Let $\mathcal{A}$ be a finite set of actions, and let $\mathcal{O}$ be a finite, exhaustive, and mutually exclusive set of outcomes.

We start with a decision making principal, who has complete and transitive preferences $\succsim$ over $\Delta(\mathcal{O})$, and n prediction making agents. We say that the principal's preferences uses some tie-breaking procedure, including over actions that produce the same distribution, so their preferences are strict. This simplifies the following theorems and proofs, although it is not crucial for the results.

The principal receives a set of predictions $p \in \Delta(\mathcal{O})^{|\mathcal{A}|\times n}$, with $p_{i,a,o}$ referring to the probability that agent i assigns to outcome o conditional on action a being chosen. Dropped subscripts indicate all values for the dropped dimension are included, and a prefacing negative sign indicates all except that subscript are included. For example, p_i refers to all predictions made by agent i and $p_{-(i,a)}$ refers to all predictions except agent i's predictions for action a. After receiving p from the agents, the principal chooses their action using a decision rule $D : \Delta(\mathcal{O})^{|\mathcal{A}|\times n} \to \mathcal{A}$.

Once action a is taken and outcome o is realized, agents are assigned scores according to a joint scoring rule $S : \mathcal{A} \times \Delta(\mathcal{O})^{|\mathcal{A}|\times n} \times \mathcal{A} \times \mathcal{O} \to \mathbb{R}^n$, with S_i referring to the function that calculates only agent i's score. For notational purposes, we let $S(a, p, o)$ refer to the vector of agent scores

after outcome o is realized, and $S(a, p, q)$ refer to the expected scores after the action is taken but before the outcome is realized, where $q \in \Delta(\mathcal{O})^{|\mathcal{A}|}$ represents the true distribution over outcomes. We use q_a to refer to the true distribution conditional on action a, and Q to refer to all agents predicting q. For now, we consider the case where the ground truth q is known to all agents.

In equilibrium, each agent chooses their report p_i to maximize their expected score S_i, conditional on each other agent's report and the decision rule D. This means that p is an equilibrium if $\nexists p'_i \in \Delta(\mathcal{O})$, such that

$$S_i(D((p'_i, p_{-i})), (p'_i, p_{-i}), q) \geq S_i(D(p), p, q)$$

A joint scoring rule/decision rule pair is *strictly proper* if there exists exactly one equilibrium, and in it each agent reports their true beliefs. Formally, $\forall q, p$, $\exists i, p'_i$ such that

$$S_i(D((p'_i, p_{-i})), (p'_i, p_{-i}), q) > S_i(D(p), p, q)$$

if and only if $p \neq Q$

Let a^* be the principal's most preferred action, so that $q_{a^*} \succ q_a,\ \forall a \in \mathcal{A} \setminus \{a^*\}$.

A joint scoring rule/decision rule pair is *jointly quasi-strictly proper* if Q is an equilibrium, in all equilibria a^* is the chosen action, and conditional on any action all agents are strictly incentivized to report honestly for that action and weakly incentivized to report honestly for all actions. Formally, $\forall\ q$, $p = Q$ is an equilibrium, $\forall p$, if $D(p) \neq a^*$, then $\exists i, p'_i$ such that

$$S_i(D((p'_i, p_{-i})), (p'_i, p_{-i}), q) > S_i(D(p), p, q)$$

and $\forall a, a'$,

$$S_i(a, (q_{a'}, p_{-(i,a')}), q) \geq S_i(a, p, q)$$

with the inequality strict if $a' = a$ and $p_{i,a} \neq q_a$.

A joint scoring rule/decision rule pair is *symmetric* if (modulo the principal's tie-breaking procedure) they remain consistent under all permutations of indices to actions, outcomes, and agents.

Theoretical Results

In the case of a single agent, (Othman and Sandholm 2010) defines the *max* decision rule to be the decision rule that chooses an action a where $p_a \succsim p_{a'}\ \forall a' \in \mathcal{A}$. A key result of their paper is that there is no symmetric scoring rule that is quasi-strictly proper when combined with the max decision rule. The practical implication of this is that there is no symmetric scoring rule/decision rule pair that deterministically chooses the principal's most preferred action.

To provide intuition,(Othman and Sandholm 2010) includes an example with are two actions, a_1 and a_2, along with two outcomes, o_1 and o_2. The principal wants to maximize the probability of o_1, and we have that $q_{a_1} = [0.5, 0.5]$ and $q_{a_2} = [0.25, 0.75]$. The agent is evaluated with the log scoring rule, which assigns a score of $\log(p_{a,o})$ when action a is chosen and outcome o is realized.

If the agent reports honestly for both actions, then action a_1 will be chosen by the max decision rule and the agent's expected score will be $0.5{*}\log(0.5){+}0.5{*}\log(0.5) = \log(0.5)$.

If the agent instead reports $p_{a_1} = [0.2, 0.8]$ while reporting honestly for a_2, then the max decision rule will select a_2 and the agent's expected score becomes $0.25\log(0.25) + 0.75\log(0.75) > \log(0.5)$. So, the agent is incentivized to misrepresent the action that the principal would take if they knew the truth, causing the principal to choose the lower variance action instead.

To begin our own contribution, we start with a couple definitions. A joint scoring rule is *linearly zero-sum, strictly proper, and symmetric (LSS)* if it has the form

$$S_i(a, p, q) = s(p_{i,a}, q_a) - \frac{\sum_{j \neq i} s(p_{j,a}, q_a)}{n-1} + h(p_{-i})$$

where h is any function and s is a symmetric, strictly proper scoring rule for the one agent and one action case. For simplicity, we assume $h(p_{-i}) = 0, \forall p_{-i}$ in all the following theorems, and we will note when that assumption is relevant.

A decision rule is *independent from suboptimal alternatives (ISA)* if, for each action, only the most preferred prediction conditional on that action is considered. Formally, if $p_{-(i,a)} = p'_{-(i,a)}$ and $\exists j$ such that $p_{j,a} \succsim p_{i,a}$ and $p_{j,a} \succ p'_{i,a}$, then $D(p) = D(p')$.

The *ISA-max* decision rule is the ISA multi-agent version of the max decision rule, choosing $D(p) = a$ if and only if $\exists i$ such that $p_{i,a} \succsim p_{j,a'},\ \forall j \in \{1, ..., n\}, a' \in \mathcal{A}$. As we will see in the following proof, the ISA-max decision rule is a good analogue for the max-decision rule when paired with an LSS scoring rule, as in all equilibria it selects the same action as applying the max-decision rule to any individual agent. We restrict our analysis to equilibrium behavior without concern about the ease of computing equilibria, since all agents reporting honestly is always an equilibrium under the mechanisms we analyze.

Before we proceed, we first introduce a useful lemma.

Lemma 1. *Under an LSS scoring rule S, all agents receive an expected score of* 0 *in any equilibrium.*

The proofs for this and all theoretical results are provided in the technical appendix.

Using this lemma, we can construct a decision rule/scoring rule pair that allows for the the Othman and Sandholm (2010) impossibility result to be overcome.

Theorem 2. *When $n \geq 2$, the combination of the ISA-max decision rule D and an LSS scoring rule S is jointly quasi-strictly proper, and in any equilibrium the max decision rule applied to any agent selects a^*.*

As this is our key contribution, and the proof gives substantial intuition, we provide it here.

Proof. First, we show that in equilibrium, $\nexists p_{i,a}$ such that $p_{i,a} \succ q_{a^*}$. Suppose p is an equilibrium, and such a prediction exists. Based on the decision rule, the principal must end up choosing some action a' where $\exists p_{j,a'} \succ q_{a^*}$. Then, since the decision rule is ISA, there exists some agent $k \neq j$ who is either already reporting honestly or can change their prediction to $p_{k,a'} = q_{a'}$ without affecting the action taken.

The score for such a prediction, $S_k(a', (q_{a'}, p_{-(k,a')}), q)$, is

$$s(q_{a'}, q_{a'}) - \frac{s(p_{j,a'}, q_{a'})}{n-1} - \frac{\sum_{i \neq j,k} S'(p_{i,a'}, q_{a'})}{n-1} > 0$$

The inequality follows because $s(\cdot, q_{a'})$ is uniquely maximized at $q_{a'}$, and $p_{j,a'} \neq q_{a'}$. By Lemma 1, this contradicts that p is an equilibrium.

Next, we show that in equilibrium, $\nexists i$ such that $q_{a^*} \succ p_{i,a^*}$. Suppose p is an equilibrium, and such a prediction exists. If another agent $j \neq i$ reports $p_{j,a^*} = q_{a^*}$, then $D(p) = a^*$ since the decision rule is ISA and we have previously established that no predictions are more preferred to q_{a^*}. The score for such a prediction, $S_k(a', (q_{a^*}, p_{-(j,a^*)}), q)$, is equal to

$$s(q_{a^*}, q_{a^*}) - \frac{s(p_{i,a'}, q_{a^*})}{n-1} - \frac{\sum_{k \neq i,j} S'(p_{k,a^*}, q_{a^*})}{n-1} > 0$$

Again by Lemma 1, this contradicts that p is an equilibrium. In equilibrium, each agent reports honestly for a^* and there are no reports $p_{i,a} \succ q_{a^*}$, so running the max decision rule on any p_i must choose a^*. Using the ISA-max decision rule across predictors similarly chooses a^*. Conditional on any action, each agent's score is given by a strictly proper scoring rule for that action's prediction, so honesty is strictly incentivized and honesty for untaken actions is weakly incentivized. As such, the decision/scoring rule pair is jointly quasi-strictly proper. □

The intuition for why this works is that from each agent's perspective, the other predictions are constant. Therefore, conditional on any action, they face a strictly proper scoring rule and are incentivized to predict honestly. Furthermore, affecting the principal's choice of action and the resulting distribution of outcomes provides no benefit to them. Since scores are zero-sum, any score increase from a distribution shift would be exactly offset by an equivalent score increase to the other agents. The ISA decision rule ensures that if agents are overstating a suboptimal action's value, then at least one agent can report honestly without changing the chosen action, and that if agents are understating the optimal action's value, then at least one agent can report honestly and ensure it is chosen.

ISA is not a necessary property for the decision rule. As we will cover in the subsection on stochastic choice, randomly choosing between which agents to believe also leads to equilibria where the principal deterministically chooses their most preferred action.

However, both ISA and stochastic choice can be difficult for the principal to commit to in the event of off-equilibrium behavior. If agents provide conflicting predictions, the principal may not be willing to throw away half the information provided. Fortunately, if we restrict to principals with preferences that follow the Independence axiom, we can use a decision rule that incorporates all predictions. The *Independence* axiom states that for any $a, b, c \in \Delta(\mathcal{O})$ and $p \in (0, 1]$, $a \succ b$ iff $pa+(1-p)c \succ pb+(1-p)c$. Preferences that meet this criteria include von Neuman-Morgenstern utility functions, as well as lexicographic preferences. The *mean-max* decision rule applies the max decision rule to the mean predictions for each action.

Theorem 3. *When $n = 2$, for a principal with preferences that follow Independence, the combination of the mean-max decision rule and an LSS scoring rule is jointly quasi-strictly proper.*

The reason for the $n = 2$ restriction is that the proof relies on any single agent being able to counteract a negative misrepresentation of a^* from other agents to ensure that the action is chosen regardless. With $n \geq 3$ agents, this is not always possible. There can be an equilibrium where $q_{a^*} \succ p_{i,a^*}$, $\forall i$, but there do not exist $\bar{p}, \underline{p}$ such that agent j predicting $p_{j,a^*} = \bar{p}$ and $p_{j,a} = \underline{p}\ \forall a \neq a^*$ results in a^* being chosen. However, if the agents are able to form coalitions and collaborate, the proof can be easily modified to work for any number of agents. If any agent is dishonest in a way that would result in a^* not being chosen, all the other agents working together can exploit that to earn a positive score.

A nice property of these jointly quasi-strictly proper mechanisms is that beyond having no incentive to lie about untaken actions, there *is* an incentive not to lie too drastically. If there exists a prediction $p_{j,a'}$ for untaken action a' for which $s(p_{j,a'}, q_{a'})$ is sufficiently low, then for $i \neq j$ there exists a prediction $p_{i,a'} \succ q_{a^*}$ that results in action a' being chosen and agent i receiving a positive expected score, along with agent j receiving a negative one.

The choice of scoring rule affects how inaccurate a prediction for an untaken action can be in equilibrium. For example, $p_{i,a',o} = 0$ when $q_{a,o} > 0$ results in an expected score of negative infinity under the log scoring rule, and only a finitely bad score under the Brier scoring rule, so whether that prediction is possible in equilibrium depends on the scoring rule chosen. The optimal choice to minimize inaccuracy depends on the principal's preferences and beliefs about the distribution of q.

Beyond merely limiting the size of divergence from honest reporting for unchosen actions, it is possible to strictly incentivize honesty for *all* actions. Consider a *disagreement-seeking-max* decision rule, which always chooses an action for which predictions differ if possible, and follows the max decision rule otherwise. Formally, if $\tilde{A}_p \equiv \{a \in \mathcal{A} | \exists i, j \ s.t. \ p_{i,a} \neq p_{j,a}\} \neq \emptyset$ then $D(p) \in \tilde{A}_p$, and otherwise $D(p) = \hat{a}$, where $p_{i,\hat{a}} \succ p_{i,a'}, \forall a' \in \mathcal{A} \setminus \{\hat{a}\}$.

Theorem 4. *If $n \geq 2$, the combination of a disagreement-seeking-max decision rule and an LSS scoring rule is strictly proper, and the principal deterministically chooses a^*.*

The downside of using a disagreement-seeking-max decision rule is that the proof for the above theorem is heavily reliant on the assumption that all agents know the ground truth q. If any of them make even slight errors, or if there is a small amount of noise in their report, then the rule instead selects a random action, with no relation to the principal's preferences. This contrasts the ISA-max and mean-max decision rules where adding a small amount of noise to either agent's prediction will for reasonable preferences still result in choosing a close to optimal action.

There are multiple decision rules that can result in a jointly quasi-strictly proper mechanism when paired with an LSS scoring rule. How important is the LSS property in the scoring rule? It can be shown that LSS scoring rules are unique

Algorithm 1: This algorithm performs a binary search over actions, using conditional predictions to narrow the action space

procedure BINARYACTIONSEARCH($\mathcal{A}, D, S$)
  while $|\mathcal{A}| > 1$ **do**
    $A_1 \leftarrow$ half of $\mathcal{A}$
    $A_2 \leftarrow$ other half of$\mathcal{A}$
    $p_{A_1}, p_{A_2} \leftarrow$ ElicitPrediction(A_1, A_2)
    $\mathcal{A} \leftarrow D((p_{A_2}, p_{A_2}))$
  end while
  return single action in A
end procedure
procedure ELICITPREDICTION(A_i, A_j)
  return predictions from n agents evaluated with S if A_i or A_j is selected
end procedure=0

in being able to achieve the incentive for honesty.

Theorem 5. *If a symmetric scoring rule can be quasi-strictly proper when paired with a decision rule for any set of complete and transitive principal's preferences, then the scoring rule is LSS.*

This is an important result, because it drastically reduces the space through which to search for the optimal scoring rule/decision rule pair. It is also highly relevant to AI alignment, as it means that incentives for honesty are unlikely to arise accidentally or by default. The explicit implementation of an LSS scoring rule is necessary for that end.

Efficient Search

If a decision problem is being used to avoid performative prediction, one danger is that the action space is not sufficiently descriptive. For example, a diner going for lunch could divide the action space into the nearby restaurants without clarifying what they would order at each one. That reintroduces the possibility of performative prediction by affecting the choice of meal, for example by influencing them to choose a cheap option by predicting they will not enjoy it. While any division of the action space reduces a prediction's influence on the outcome, ideally we would like that influence to be as small as possible. To that end, we could end up with an enormous action space, and very small details distinguishing different actions. Fortunately, even a large action space can be searched efficiently.

Theorem 6. *A principal can identify a^* with at most $O(log(|\mathcal{A}|))$ comparisons between actions.*

The proof shows that Algorithm 1, which essentially does a binary search through action space, maintains the incentives for honesty even as predictions can affect later predictions. Here, we rely on setting the term $h(p_{-i})$ of the LSS scoring rule to zero for all p_{-i}.

One way of implementing Algorithm 1 would be to add more and more details to the action with each iteration. For example, when deciding on lunch plans, one could first ask what cuisine they would most enjoy, then what restaurant, then which menu item, accumulating additional information with each step. Iteratively eliciting information is not necessary, however, and using powerful predictive models it is possible to jump straight to the end of the process.

Theorem 7. *A principal can identify a^* with at most $O(1)$ comparisons between actions.*

The way this works is by eliciting a non-conditional, non-zero-sum prediction from an agent about which action will be chosen, then running Algorithm 1 but with the action space instead split into the set containing only that action and the set containing all other actions. a^* is then identified after a single comparison.

Stochastic Choice

The principal may be interested not only in identifying a^*, but in the information contained in predictions about untaken actions. For example, consider training an AI system that generates potential actions in the way LLMs generate potential text completions. We would like to have a reward model evaluate the outcomes of each action, but in an online training environment only one can be taken, so we instead evaluate the predicted outcomes. Performative prediction here would be disastrous, as it would train the model to generate actions that are easy to predict rather than actions that the developers desire, but even a jointly quasi-strictly proper scoring rule is insufficent. We would also want accurate predictions for the untaken actions, so they can be properly evaluated.

If the principle is willing to partially randomize their decision, then for reasonable methods of randomization, the actions for which honest predictions are strictly incentivized can be greatly expanded. This can be done with arbitrarily small probability of not taking a^*, and with zero chance of taking deeply unpalatable actions.

Let $D : \Delta(\mathcal{O})^{|\mathcal{A}|\times n} \rightarrow \Delta(\mathcal{A})$ be a non-deterministic decision function, with D_a representing the probability assigned to action a. Consider the following regularity conditions:

Condition 1. *If $p_{-(i,a)} = p'_{-(i,a)}$ and $p'_{i,a} \succ p_{i,a}$ then $D_a(p) > 0$ implies $D_a(p') > 0$*

Condition 2. *If $p_{-(i,a)} = p'_{-(i,a)}$ and $p_{i,a} \succ p'_{i,a}$ then for $a' \neq a$ $D_{a'}(p) > 0$ implies $D_{a'}(p') > 0$*

Condition 1 says that if an agent changes their prediction for an action to a more preferred distribution, this will not cause that action to be assigned zero probability. Condition 2 says that if an agent changes their prediction for an action to a less preferred distribution, this will not cause other actions to be assigned zero probability.

These conditions ensure that in equilibrium, $S_i(a, p, q) = 0$, $\forall a$ s.t. $D_a(p) > 0$. This is distinct from Lemma 1, which only showed that $E_{a\sim D(p)}[S_i(a, p, q)] = 0$ in equilibrium.

Lemma 8. *Under an LSS scoring rule S and ISA decision rule D, if Conditions 1 and 2 are met then in equilibrium p $\forall i, \forall a$ such that $D_a(p) > 0$, $p_{i,a} = q_a$*

In practice, we may be willing to accept some chance of taking non-optimal actions in order to gather information from the predictions for these actions. While Chen et al. (2011) showed that randomizing over actions with full support can

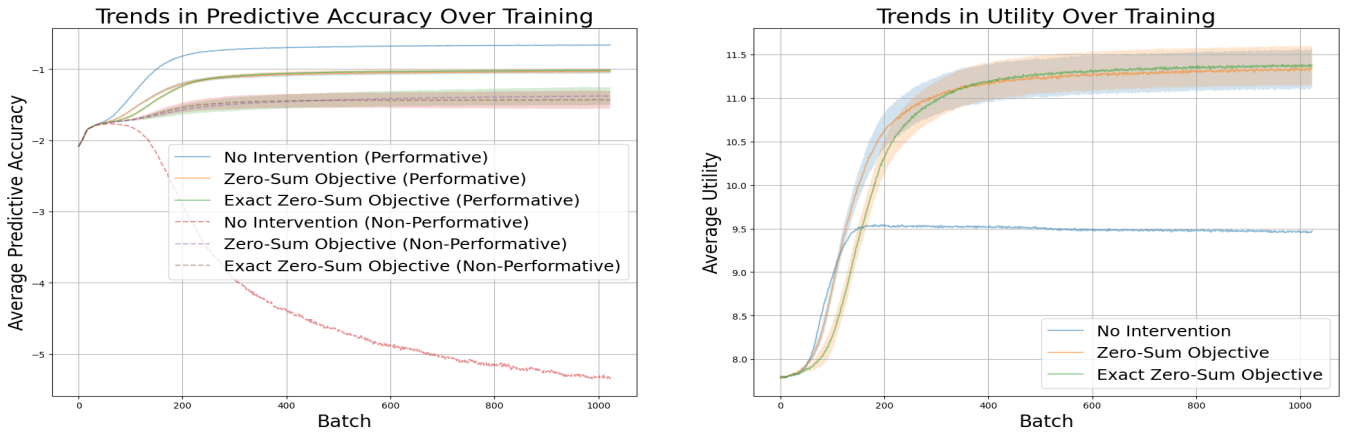

Figure 1: (Left) In environments that incentivize performative prediction, training with an LSS objective avoids the model becoming performative, and results increasing accuracy across predictions. Models trained with no intervention are more accurate for whichever action is chosen, as they influence the choice to be easier to predict, but are less accurate overall. Error bars show that the difference between either zero-sum intervention and the no intervention baseline is statistically significant. (Right) When no intervention prevents a model from becoming performative, user utility plateaus earlier and at a lower level. Error bars indicate that the same statistical significances apply to utility.

incentivize predictions for all actions, we show that with multiple agents we can randomize with only partial support and still strictly incentivize honest predictions for those actions. In particular, the principal can restrict support to only the actions they would be willing to randomize over if they knew the true distribution. This means they do not need to commit to or accept taking to catastrophically bad actions with any probability.

To do so, we require one more condition:

Condition 3. *If* $p_{-(i,a)} = p'_{-(i,a)}$ *and* $D_a(p) = D_a(p') = 0$ *then* $D(p) = D(p')$

This says that if an agent modifies their prediction for an action without changing that it is assigned zero probability, the probabilities assigned to other actions do not change.

Theorem 9. *Under an LSS scoring rule* S *and ISA decision rule* D*, if Conditions 1- 3 are met then in any equilibrium* p*,* $D(p) = D(Q)$.

This result is particularly useful if we want to train a model to take actions that are predicted to have good outcomes. It allows us to strictly incentivize honesty for relevant untaken actions, while ensuring safe exploration. Knowing that some actions are too bad to be considered is also useful, even without being guaranteed honest predictions for them.

Stochastic choice can also be used to deterministically take a^*, with needing a decision rule that satisfies ISA or independence. We call the decision rule that randomizes which agent to believe and then takes their most preferred prediction the *random-max* decision rule.

Theorem 10. *When* $n \geq 2$*, an LSS scoring rule and the random-max decision rule is jointly quasi-strictly proper.*

Experiments

We test our main theoretical result in a toy environment, with eight possible actions, eight possible outcomes, and eight variables representing context. The ground truth probabilities are given by a randomly initialized neural net that takes in as input the context and a one-hot vector representing the choice of action, and outputs a distribution over outcomes. A principal with a randomly generated utility function over actions makes their decisions by taking the softmax of their expected utility from each action, consolidating reports from agents in an ISA manner.

We train models to predict the outcome, using a cross-entropy loss function and running the gradient through the impact of the prediction on the principal's decision. This is the simplest way to implement performativity in a toy environment, and shows the robustness of zero-sum competition in avoiding it.

Our first experiment compares training with no intervention to two methods of implementing an LSS objective. The first method trains an agent against a detached version of itself that makes identical predictions, which we call *exact*. The second uses dropout to generate two different predictions from the same model, then performs a gradient update for each one while detaching the other. This provides evidence on behavior when agents have different information and thus disagree about probabilities. We apply dropout to all models in order to isolate the effect of an LSS objective.

In Figure 1 we compare how the various training methods affect predictive accuracy, both when taking performativity into account and when weighing the predictions for all actions equally. We also measure how the principal's utility changes throughout the training process.

We can see that both implementations of an LSS objective perform very similarly, increasing in both measures of predictive accuracy. Performative predictive accuracy is slightly higher, as higher utility actions tend to have more extreme distributions, resulting in a higher prediction score. In contrast, training without an intervention leads to the

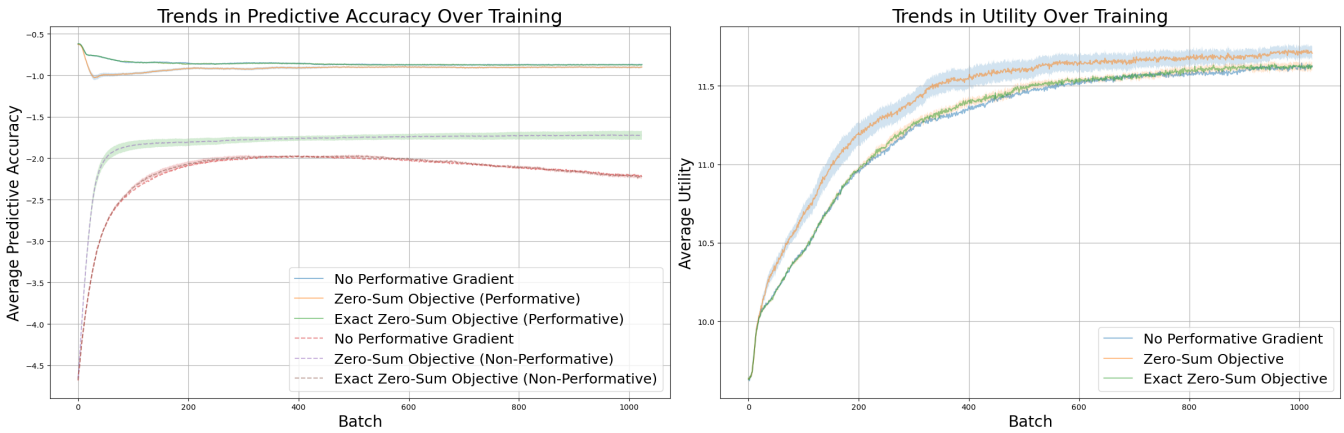

Figure 2: (Left) An LSS objective using two different predictions from the same model results in faster and larger decreases in performativity than an exact LSS objective or training in a non-performative environment. Error bars show that the difference between the inexact zero-sum objective and the non-performative training baseline is statistically significant, while the difference from the exact zero-sum objective is not. (Right) The decrease in performativity leads to higher utility for users, with larger gains for larger drops in performativity. Error bars indicate that the same statistical significances apply to utility.

largest gain in performative predictive accuracy, but after an initial increase non-performative predictive accuracy sharply drops. When this divergence occurs, the principal's utility plateaus, while for LSS objectives it continues to rise. Without an intervention, performativity compounds throughout training. The more inaccurate the conditional prediction for an action is, the more the local gradient pushes towards performativity to ensure that action is not taken. Similarly, it discourages gradient updates from making the prediction more accurate if doing so increases the chance the action will be chosen.

Our second experiment tests whether a model that has already become performative can have that behavior trained out of it. We compare the same two implementations of an LSS objective, alongside removing the gradient that runs through the principal's choice of action. This last intervention can be thought of as training to predict only historical data, rather than making predictions that can affect their own outcome.

Figure 2 compares the different training methods using the same measures as in the first experiment. We can see that the exact LSS objective behaves like training in a non-performative environment, which makes sense since they produce nearly identical gradients. The LSS objective that generates two distinct predictions untrains performativity faster, plateaus at a higher level of predictive accuracy, and results in higher utility for the principal. We speculate that this results slight differences in predictions allow the gradient to get un-stuck by providing the more accurate agent an incentive to have that action be chosen.

We further run robustness checks to ensure that the results are not affected by experimental choices. No major changes were observed after changing the decision rule from ISA to mean, only assigning positive probability to above-median expected utility actions, changing the scoring rule base from log score to Brier score, sampling more than two agents when calculating the LSS objective, or pretraining the model on historical data. These results are available in the technical appendix, along with further experimental details.

Discussion and Future Direction

Our work demonstrates that it is possible to incentivize honesty in conditional predictions, both theoretically and practically. This allows for the elicitation of conditional predictions as a safe alternative to accepting performativity in the unconditional case.

Eliminating the risk of performative prediction from powerful AI systems represents an important step towards their safe usage. This applies both to Oracle AI for aligning superintelligent systems, and the use of predictive models for either economic purposes or training of more powerful systems. Honesty means we can trust the predictions they output, putting the onus on us to use that information wisely.

We are interested in both empirical and theoretical expansions of this work. As we have only demonstrated the success of our intervention in a toy environment, it would be useful to extend it to a more complex environment, especially one using real world data. It would also be valuable for theoretical progress loosening the assumption that all agents have the same information. While this is less concerning when both agents are contained in the same model, and our experiments showed this does not result in dishonesty, it would provide additional robustness. Here, we may be interested in the regret dynamics as agents learn each others' information, such as the untruthful swap regret (Fujii 2023). Once we can conclusively eliminate concerns about performative prediction, we can move on to the next question: how can we use honest predictions to align powerful AI systems and ensure a safer world?

Acknowledgements

We thank Gabriel Carroll, Evan Hubinger, Johannes Treutlein, Dan Valentine, and Simon Marshall for valuable conversations and feedback on this paper. We also thank three anonymous reviewers for their helpful comments.

References

Armstrong, S. 2013. Risks and Mitigation Strategies for Oracle AI. In *Philosophy and Theory of Artificial Intelligence*, 335–347. Springer.

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

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

Bell, J.; Linsefors, L.; Oesterheld, C.; and Skalse, J. 2021. Reinforcement Learning in Newcomblike Environments. *NeurIPS*, 35.

Bengs, V.; Hüllermeier, E.; and Waegeman, W. 2023. On Second-Order Scoring Rules for Epistemic Uncertainty Quantification. *arXiv preprint arXiv:2301.12736*.

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

Brier, G. W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1).

Chan, A. 2022. Scoring Rules for Performative Binary Prediction. *arXiv preprint arXiv:2207.02847*.

Chen, Y.; Kash, I.; Ruberry, M.; and Shnayder, V. 2011. Decision markets with good incentives. In *International Workshop on Internet and Network Economics*, 72–83. Springer.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *NeurIPS*, 31: 22146–22157.

De-Arteaga, M.; and Elmer, J. 2022. Self-fulfilling prophecies and machine learning in resuscitation science. *Resuscitation*.

Fujii, K. 2023. Bayes correlated equilibria and no-regret dynamics. *arXiv preprint arXiv:2304.05005*.

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

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

Hadfield-Menell, D.; and Hadfield, G. 2019. Incomplete Contracting and AI Alignment.

Hanson, R. 2003. Combinatorial Information Market Design. *Information Systems Frontiers*, 5(1): 107–119.

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

Hendrycks, D. 2023. Natural Selection Favors AIs over Humans. arXiv:2303.16200.

Hubinger, E.; Jermyn, A.; Treutlein, J.; Hudson, R.; and Woolverton, K. 2023. Conditioning Predictive Models: Risks and Strategies. *arXiv preprint arXiv:2302.00805*.

Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; and Garrabrant, S. 2019. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

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

Krueger, D.; Maharaj, T.; and Leike, J. 2020. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*.

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

Oesterheld, C.; and Conitzer, V. 2020. Minimum-regret contracts for principal-expert problems. *Conference on Web and Internet Economics (WINE)*, 16.

Oesterheld, C.; Treutlein, J.; Cooper, E.; and Hudson, R. 2023. Incentivizing honest performative predictions with proper scoring rules.

Omohundro, S. M. 2008. The Basic AI Drives. In *Proceedings of the 2008 conference on Artificial General Intelligence: Proceedings of the First AGI Conference*, 483–492. IOS Press.

Othman, A.; and Sandholm, T. 2010. 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*, 625–632. Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS).

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

Savage, L. J. 1971. Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*, 66: 783–801.

Shah, R.; Varma, V.; Kumar, R.; Phuong, M.; Krakovna, V.; Uesato, J.; and Kenton, Z. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. *arXiv preprint arXiv:2210.01790*.

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

Turner, A.; Smith, L.; Shah, R.; Critch, A.; and Tadepalli, P. 2021. Optimal Policies Tend To Seek Power. *NeurIPS*, 34: 23063–23074.