Decision-Relevant Predictions with Joint Scoring Rules

Anonymous Authors[1]

Abstract

When a decision-making principal wishes to elicit predictions from an agent using a proper scoring rule, the principal must specify the outcomes to be assigned probability. This can result in them either missing important considerations or facing too much information to process. We would instead like the agent making the prediction to provide the decision-relevant set of outcomes. This could be either the simplest set of outcomes leading the principal to the decision they would make under full information, or the simplest set of outcomes where no further information would change the principal's decision. We show that it is impossible to strictly incentivize an agent to provide these sets, but that it can be done with two or more agents evaluated jointly. This is done by rewarding agents refining the outcome set in a way that changes the principal's decision. We then show how to adapt the mechanism to elicit a prediction in simplified form, for cases where the set of possible outcomes is very large or continuous. These mechanisms can be used to augment human judgment, and in the AI training process for auditing and aligning their decision making.

1. Introduction

While making a decision among possible courses of action, a common step is to predict the likely outcomes under each. For important decisions outside of our normal expertise, we may consult an expert and use their predictions instead of or in addition to our own. While it is possible to incentivize honest predictions from an expert using a proper scoring rule (Brier, 1950; Good, 1952), there is no best practice for determining which outcomes to predict over.

A decision maker who constructs an outcome set can miss important considerations, leading to suboptimal decisions even after eliciting an expert's prediction. However, expanding the outcome set to cover all possible details can quickly make predictions too large to evaluate, or even elicit. What we would like instead is for the expert, who knows which outcomes are relevant, to partition the set of outcomes so that the decision maker gets all necessary information but no extraneous details.

The challenge with eliciting partitions is assessing their relevance. Only one action can be implemented, so the decision maker cannot tell if they missed important information on untaken actions. This issue arises even when experts are divided into separate partition and prediction roles.

Our results show that while a single agent cannot be incentivized to provide the simplest grouping that leads to the best action, using two agents can achieve this outcome. To do so, once the first agent has provided a partition of outcomes, the other agent can provide a 'second opinion' that further splits outcomes. If this changes the principal's choice of action, the second agent is rewarded and the first is punished. The first agent is then incentivized to simplify the outcome space only insofar as it does not change the principal's decision. We can elicit further simplification while still choosing the optimal action with a mechanism that allows the first agent to reply, and either mechanism can be modified to work with a single type of agent if they are memoryless, like AI.

We then examine the case where we wish to have agents provide both a simplified partition and prediction. This can be done by allowing the second agent to either split outcomes or provide an alternative prediction, which keeps the first agent honest on both fronts.

1.1. Literature Review

The literature on eliciting predictions over a set of outcomes is extensive, starting with Brier (1950) and Good (1952). Gneiting & Raftery (2007) establishes properties that apply to all (strictly) proper scoring rules, which are defined as those that (strictly) incentivize reporting actual beliefs. Blackwell (1951) compares information structures on their ability to lead to better choices on all decision problems, while we focus on some specific decision problem.

The closest paper to our work is Lipnowski et al. (2020), which models a principal without attention cost summariz-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ing information for a decision making agent with attention costs. They focus on the case that the principal and the agent share the same utility function, while our interest is in adapting scoring rules.

Eliciting conditional predictions for use in making decisions was investigated by Othman & Sandholm (2010), who showed that it is impossible to use predictions from a single agent to deterministically take the best action available. Follow-up work (Chen et al., 2011; Oesterheld & Conitzer, 2020) found partial workarounds in the single agent case, and Hudson (2025) showed that multiple agents in a zero-sum competition can circumvent the impossibility result. These papers focus on eliciting accurate predictions, rather than decision-relevant ones.

The mechanism of rewarding agents for changing the principal's mind appeared in Grace (2014), which suggested that applying it repeatedly would converge to the truth. Debate between two agents under the same assumption that the truth will eventually converge was explored by Irving et al. (2018) and Brown-Cohen et al. (2025). Our work differs in that outputs are restricted to be partitions, limiting the opportunity for deceptive arguments.

2. Model

Let $\mathcal{A}$ be a finite set of actions, and let Ω be a set of outcomes. A partition Π of a set X is a finite collection $Y = \{y_1, \ldots, y_k\}$ of nonempty, pairwise disjoint subsets with $\bigcup_{j=1}^{k} y_j = X$. Given two partitions of X, denoted Π_1 and Π_2, we say Π_2 is a *coarsening* of Π_1 (and Π_1 is a *refinement* of Π_2) if every element of Π_1 is wholly contained in some element of Π_2. A coarsening (resp. refinement) is *strict* if $|\Pi_2| < |\Pi_1|$ (resp. $|\Pi_2| > |\Pi_1|$).

Let $\mathcal{C}_\Pi$ denote all coarsenings of Π and $\mathcal{C}_\Pi^>$ strict coarsenings, with $\mathcal{R}_\Pi$ and $\mathcal{R}_\Pi^>$ the equivalent of refinements. Let $\mathcal{C}_\Omega$ denote the set of all partitions of Ω.

Given a partition Π of Ω, a set of conditional predictions $P \in (\Delta\Pi)^{\mathcal{A}}$ consists of a prediction $p_a \in \Delta\Pi$ for each $a \in \mathcal{A}$. The probability assigned to outcome $\pi \in \Pi$ conditional on action a is denoted $p_{a,\pi}$. For $\Pi_C \in \mathcal{C}_\Pi$, and prediction P on Π we let P_{Π_C} denote the set of predictions over Π_C where $p_{a,\pi_C} = \sum_{\pi \subset \pi_C} p_{a,\pi}$.

We have two agents, denoted 1 and 2, who are risk-neutral, and only interested in a reward that can be costlessly provided by the principal. Both agents have beliefs $\mu \in (\Delta\Omega)^{\mathcal{A}}$ regarding the distribution of outcomes, with μ_a denoting the distribution conditional on action a. When agent i reports a partition, it is denoted Π_i, and when they report a prediction, it is denoted P^i.

The *simplicity* of a partition $\Pi \in \mathcal{C}_\Omega$ is measured by a simplicity function $S : \mathcal{C}_\Omega \to \mathbb{R}$ that is strictly monotone with coarsening: if $\Pi_2 \in \mathcal{C}_{\Pi_1}$ then $S(\Pi_2) \geq S(\Pi_1)$, with strict inequality when $\Pi_2 \in \mathcal{C}_{\Pi_1}^>$. Two examples are (i) cardinality-based $S_{\text{card}}(\Pi) = -|\Pi|$, and (ii) entropy-based $S_{\text{ent}}^P(\Pi) = \sum_{a \in \mathcal{A}} \sum_{\pi \in \Pi} p_{a,\pi}^{\Pi} \log p_{a,\pi}^{\Pi}$. The latter measures the simplicity of a partition when it is applied to a particular prediction, P. Without loss of generality, we let S have range $[0, 1]$, which can be accomplished by applying the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$.

We also have a distance metric between actions, $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, and a threshold $\delta > 0$ for how different actions must be for that difference to be relevant. This can be the discrete metric, where $d(x, y) = 0$ if $x = y$ and 1 otherwise.

Finally, we have a decision-making principal, who uses a rule $D : \bigcup_{\Pi \in \mathcal{C}_\Omega} (\Delta\Pi)^{\mathcal{A}} \to \mathcal{A}$ to map conditional predictions to actions. Agents are assumed to know the decision rule, since the principal will not want to conceal it.

The principal may either want agents to provide the simplest partition such that no further refinements change would change their choice of action by more than δ, or the simplest partition such that they choose an action within δ of what they would choose under full information. The set of all partitions meeting the first criteria is defined as

$$\Pi^* = \arg\max_{\Pi \in \mathcal{R}_\delta^S} S(\Pi)$$

where

$$\mathcal{R}_\delta^S = \{\Pi | d(D(P_\Pi), D(P_{\Pi'})) < \delta, \forall\, \Pi' \in \mathcal{R}_\Pi\}$$

while the set of all partitions meeting the second criteria is

$$\Pi^{*+} = \arg\max_{\Pi \in \mathcal{R}_{final}^S} S(\Pi)$$

where

$$\mathcal{R}_{final}^S = \{\Pi | d(D(P_\Pi), D(P_{\Pi_\Omega})) < \delta\}$$

3. Results

3.1. Simplifying Predictions

We first analyze the case where an initial expert prediction P_Ω on a finite partition Π_Ω is given. Both agents know it, and the principal can access it, but determining $D(P_\Omega)$ is expensive in cognitive effort. This corresponds to the case where the agents are independent heads on a neural network with a prediction-making head, and so have access to the same information.

The principal cannot in general strictly incentivize a single agent to report a partition in Π^*.

Theorem 3.1. *There does not exist a partition scoring rule* $\mathbb{S} : \mathcal{C}_\Omega \to \mathbb{R}$ *such that* $\arg\max_{\Pi \in \mathcal{C}_\Omega} \mathbb{S}(\Pi) \in \Pi^*$ *for all* P

Proofs are provided in Appendix A. The intuition is that if the maximal coarsening does not receive the highest score among partitions, it will not get reported when it is in Π^*, but if it does then an agent can report it without the decision maker realizing that they are missing information.

To elicit a partition in Π^*, consider the following setup, which we call the *second opinion mechanism*. First, agent 1 provides a partition $\Pi_1 \in \mathcal{C}_{\Pi_\Omega}$. Then, agent 2 provides a refinement, $\Pi_2 \in \mathcal{R}_{\Pi_1} \cap \mathcal{C}_{\Pi_\Omega}$. Rewards are

$$\begin{cases} (S(\Pi_1), S(\Pi_2) - S(\Pi_1)) & \text{if } d(D(P_{\Pi_1}), D(P_{\Pi_2})) < \delta \\ (S(\Pi_1) - 2, S(\Pi_2) - S(\Pi_1) + 2) & \text{otherwise} \end{cases}$$

The number 2 is larger than any $S(\Pi)$, which is restricted to $[0, 1]$. The principal's actual decision does not affect the agent incentives, so we can leave their choice when $D(P_{\Pi_1}) \neq D(P_{\Pi_2})$ undetermined.

Theorem 3.2. *In the second opinion mechanism, a subgame perfect equilibrium (SPE) exists, and in any SPE* $\Pi_2 = \Pi_1 \in \Pi^*$.

The intuition here is that the first agent is rewarded proportionally to the simplicity of their partition, but if there exists a refinement that changes the decision the second agent is incentivized to provide it and the first agent will lose more than they can gain from simplicity alone.

To elicit a partition in Π^{*+}, consider the following setup, which we call the *rebuttal mechanism*. First, agent 1 provides a partition $\Pi_1 \in \mathcal{C}_\Omega$. Then agent 2 provides a refinement $\Pi_2 \in \mathcal{R}_{\Pi_1} \cap \mathcal{C}_{\Pi_\Omega}$. Finally, agent 1 provides a refinement $\Pi_3 \in \mathcal{R}_{\Pi_2} \cap \mathcal{C}_{\Pi_\Omega}$. Rewards are

$$\begin{cases} (\frac{S(\Pi_1)+S(\Pi_3)}{2}, S(\Pi_2) - S(\Pi_1)) \\ \qquad \text{if } d(D(P_{\Pi_3}), D(P_{\Pi_1})) < \delta \\ (\frac{S(\Pi_1)+S(\Pi_3)}{2} - 2, S(\Pi_2) - S(\Pi_1) + 2) \\ \qquad \text{otherwise} \end{cases}$$

Theorem 3.3. *In the rebuttal mechanism, an SPE exists, and in any SPE* $\Pi_3 = \Pi_2 = \Pi_1 \in \Pi^{*+}$.

In contrast to the second-opinion mechanism, the second agent is not incentivized to provide a refinement that changes the decision if a further refinement from the first agent can change it back. In that case, the second agent would not receive the bonus for changing the decision, but would lower their score by providing a less simple partition.

3.2. Memoryless Elicitation

An advantage of using an AI to identify decision-relevant outcomes is that they are memoryless, so a single system can assume the roles of both experts, without introducing incentives for collusion. To train an AI to do this, we would like to have all agents take the same kind of inputs, produce the same kind of outputs, and be evaluated according to the same reward function.

The necessary incentives are provided through the following setup. We start by letting Π_0 to be the maximal coarsening, a single outcome occurring with probability 1. Then, we repeatedly show agent i the partition Π_{i-1} and asked it to produce a refinement $\Pi_i \in \mathcal{R}_{\Pi_{i-1}}$. If $\Pi_{i-1} = \Pi_i$, the process ends, otherwise it proceeds to agent $i + 1$.

To elicit a partition in Π^*, we assign agent i the following score, creating the *memoryless second opinion mechanism*:

$$S(\Pi_i) - S(\Pi_{i-1}) + \mathbb{1}_{i-1,i}[1 + S(\Pi_{i-1}) - \mathbb{1}_{i,i+1}[(1 + S(\Pi_{i+1}))]]$$

where $\mathbb{1}_{i,j}$ is an indicator variable for $d(D(P_{\Pi_i}), D(P_{\Pi_j})) \geq \delta$.

If an agent's partition changes the decision from the previous partition, it receives a bonus of $1+S(\Pi_{i-1})$, and is therefore incentivized to do so if possible. Even if the subsequent partition changes the decision back, the agent's score will end up higher than if they had provided a simpler partition that did not change the decision.

Under the memoryless second opinion mechanism, when $d(D(P_{\Pi_i}), D(P_{\Pi_{i+1}})) \geq \delta$, agent i has their score drop by $S(\Pi_{i+1})$, while agent $i + 1$ has their score increase by $S(\Pi_i)$, for a total score increase of $S(\Pi_i) - S(\Pi_{i+1})$. To prevent certain forms of potential collusion between agents, the mechanism can be made zero-sum by subtracting this amount from either agent $i + 2$ or the final agent.

To elicit a partition in Π^{*+}, we assign agent i the following score, creating the *memoryless rebuttal mechanism*:

$$S(\Pi_i) - S(\Pi_{i-1}) + \mathbb{1}_{i-1,i}[1 - \mathbb{1}_{i,i+1}]$$

In contrast to the memoryless second opinion mechanism, if an agent's partition changes the decision but the subsequent partition changes it again, the bonuses cancel out and they are left with a lower score than simply repeating the previous partition. Due to the lack of memory, agent i is not incentivized to change the decision back to $D(P_{i-2})$, only to an action that is different from $D(P_{i-1})$.

The difference between these mechanisms is that the memoryless second opinion mechanism rewards an additional $S(\Pi_{i-1})$ for changing the decision, and subtracts an additional $S(\Pi_{i+1})$ if the following agent changes the decision . If both occur, the agent receives a score of $S(\Pi_i) - S(\Pi_{i+1})$, greater than the 0 they would receive for reporting $\Pi_i = \Pi_{i-1}$. Under the memoryless rebuttal mechanism, an agent who changes the decision and then has it changed gets score $S(\Pi_i) - S(\Pi_{i-1}) < 0$. So, agents will anticipate that subsequent agents will change the decision whenever possible under the memoryless second opinion mechanism, but only

when it cannot be changed further under the memoryless rebuttal mechanism.

Theorem 3.4. *In the memoryless second opinion mechanism, an SPE exists, and in any SPE* $\Pi_2 = \Pi_1 \in \Pi^*$. *In the memoryless rebuttal mechanism, an SPE exists, and in any SPE* $\Pi_2 = \Pi_1 \in \Pi^{*+}$.

3.3. Eliciting Simplified Predictions

So far, we have been assuming that an initial expert prediction P_Ω on a partition Π_Ω is given. However, predictions on Ω may not be provided, or even providable, due to their complexity. Instead, we would like to directly elicit predictions in a simplified form.

When eliciting conditional predictions, those conditioned on untaken actions cannot be evaluated for accuracy. Hudson (2025) showed that when both agents have the same beliefs, it is possible for the combination of a joint scoring rule and decision rule to be jointly quasi-strictly proper, meaning that in every equilibrium the principal's preferred action, denoted a^* is chosen, agents are strictly incentivized to report truthfully for the chosen action, and weakly incentivized to report truthfully for unchosen actions.

To do this, after taking action a and observing outcome ω the score for agent $i \neq j$ is given by $\mathcal{S}_i(P^i, P^j, \omega) = s(p^i_a, \omega) - s(p^j_a, \omega)$, where s is a strictly proper scoring rule, P^i and P^j are the sets of predictions made by agents i and j respectively. The principal takes the action corresponding to the most preferred prediction made by either agent, even if the other agent disagrees for that action.

Our mechanisms can then be combined with a sequential version of the Hudson (2025) mechanism to elicit quasi-strictly proper predictions over a partition in Π^* or Π^{*+}.

This requires two conditions, the first of which is that the simplicity measure cannot depend on the predicted distributions for untaken actions, as this would violate the weak incentive for honesty on untaken actions. The cardinality metric meets this condition.

Condition 1. If $p_{D(P),\Pi_1} = p'_{D(P'),\Pi_1}$ then $S^P(\Pi_1) = S^{P'}(\Pi_1)$ for all P, P'.

The second condition is that iteratively providing strict refinements on a partition causes the decision to converge.

Condition 2. For any prediction P and sequence of partitions $\{\Pi_k\}_{k=0}^{\infty}$ where for all n, $\Pi_{n+1} \in \mathcal{R}_{\Pi_n}$, there exists some $N \in \mathrm{N}$ such that for all $n, m \geq N$, $d(D(P_{\Pi_m}), D(P_{\Pi_n})) \leq \delta$

To elicit simplified predictions, the *zero-sum second opinion* mechanism works as follows. First, agent 1 reports a partition, Π_1 and an associated prediction $P^1_{\Pi_1}$. Agent 2 can then either report the same partition and make a prediction $P^2_{\Pi_1}$ over the same partition, or suggest a strict refinement $\Pi_2 \in \mathcal{R}^{>}_{\Pi_1}$ and make a prediction $P^2_{\Pi_2}$ over it. If agent 2 provides a partition, agent 1 then makes a prediction over it, $P^1_{\Pi_2}$. Once both agents have made predictions for the same partition, the principal chooses the action corresponding to the prediction they most prefer across both agents, denoted $\mathcal{D}(P^1, P^2)$.

The scores for the agents are broken down into a second opinion component and a prediction component. The former is the score the agents would receive from the second opinion mechanism, replacing $D(P_{\Pi_2})$ with $\mathcal{D}(P^1_{\Pi_2}, P^2_{\Pi_2})$, and treating the case where the second agent responds with only a prediction as reporting $\Pi_2 = \Pi_1$.

The prediction component of the score is for agent i is $\alpha_i(P^i, P^j)(s(P^i, \omega) - s(P^j, \omega))$, where s is a strictly proper scoring rule and both P^i and P^j are made over the same partition. α_i scales the score so that the expected loss from making a dishonest prediction while the other agent predicts honestly is greater than any gain from changing the apparent distance between actions.

$$\alpha_i(P^i, P^j) = \begin{cases} 1 & \Pi_1 = \Pi_2 \\ 0 & \Pi_1 \neq \Pi_2, p^i_{\tilde{a}} = p^j_{\tilde{a}} \\ \frac{2}{E_{\omega \sim p^j_{\tilde{a}}}[s(p^i_{\tilde{a}}, \omega) - s(p^j_{\tilde{a}}, \omega)]} + 1 & otherwise \end{cases}$$

where $\tilde{a} = \mathcal{D}(P^i, P^j)$

When the above conditions hold, this mechanism is quasi-strictly proper and incentivizes reporting partitions in Π^*.

Theorem 3.5. *For the zero-sum second opinion mechanism under Conditions 1 and 2, an SPE exists, and in any SPE agent* 1 *reports some* $\Pi_1 \in \Pi^*$, *agent* 2 *responds with a prediction,* $\mathcal{D}(P^1_{\Pi_1}, P^2_{\Pi_1}) = a^*_{\Pi_1}$, *and both agents are strictly incentivized to predict honestly for* $a^*_{\Pi_1}$ *and weakly incentivized to predict honestly for all actions.*

4. Applications and Discussion

These results can be used to augment human decision making with AI. Eliciting decision-relevant predictions from AI can provide additional information while simultaneously reducing attention costs. AI forecasting is quickly approaching the level of top human experts, but even in areas where it lags it is straightforward for a human principal to further refine partitions provided by AI agents.

In addition to human principals, decision-relevant predictions can be elicited for AI principals, helping humans to audit the AI's decision-making process. The optimal partitions for different principles can be combined to highlight differences in values, which can then be used for AI control and alignment. Appendix B goes into further detail on how the mechanisms in this paper can be usefully applied.

Acknowledgments

We are thankful to various people/groups, to be named after double-blind review.

Impact Statement

References

Blackwell, D. Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1:93–102, 1951.

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

Brown-Cohen, J., Irving, G., and Piliouras, G. Avoiding obfuscation with prover-estimator debate, 2025. URL https://arxiv.org/abs/2506.13609.

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

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

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

Grace, K. How to buy a truth from a liar. https://meteuphoric.com/2014/07/21/how-to-buy-a-truth-from-a-liar/, 2014. Blog post.

Hudson, R. Joint scoring rules: Competition between agents avoids performative prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26): 27339–27346, Apr. 2025. doi: 10.1609/aaai.v39i26.34944. URL https://ojs.aaai.org/index.php/AAAI/article/view/34944.

Irving, G., Christiano, P., and Amodei, D. Ai safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

Lipnowski, E., Mathevet, L., and Wei, D. Attention management. *AER: Insights*, 2(1):17–32, 2020. doi: 10.1257/aeri.20190165.

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

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

5. Appendix A: Proofs

Theorem 5.1. *There does not exist a partition scoring rule* $\mathbb{S} : \mathcal{C}_\Omega \to \mathbb{R}$ *such that* $\arg\max_{\Pi \in \mathcal{C}_\Omega} \mathbb{S}(\Pi) \in \Pi^*$ *for all* P

Proof. Let $\mathbb{S}$ be a partition scoring rule and let $\Pi_0 = \{\Omega\}$ be the maximal coarsening.

If $\Pi_0 \in \arg\max_{\Pi \in C_\Omega} \mathbb{S}(\Pi)$, consider P_1 where $d(D((P_1)_{\Pi_0}), D(P_1)) \geq \delta$. $\Pi_0 \notin R^S_\delta$ for P_1, and hence $\Pi_0 \notin \Pi^*(P_1)$. Since $\Pi_0 \in \arg\max S$, the optimal partition is not in $\Pi^*(P_1)$.

If $\Pi_0 \notin \arg\max_{\Pi \in \mathcal{C}_\Omega} \mathbb{S}(\Pi)$ then for any $\Pi' \in \arg\max S$, $\Pi' \neq \Pi_0$. Consider P_2 such that $\Pi_0 \in R^S_\delta$. By the strict monotonicity of the simplicity function with coarsening, Π_0 is the unique simplicity-maximizer over R^S_δ, giving $\Pi^*(P_2) = \{\Pi_0\}$. Since $\Pi' \in \arg\max S$ and $\Pi' \neq \Pi_0$, the optimal partition is not in $\Pi^*(P_2)$.

As such, $\arg\max_{\Pi \in \mathcal{C}_\Omega} \mathbb{S}(\Pi) \not\subseteq \Pi^*$ for some P for all $\mathbb{S}$. □

Theorem 5.2. *In the second opinion mechanism, a subgame perfect equilibrium (SPE) exists, and in any SPE* $\Pi_2 = \Pi_1 \in \Pi^*$.

Proof. We prove this with backward induction over the game tree. Given Π_1, the only way for the second agent to achieve their highest score of $S(\Pi_2) - S(\Pi_1) + 2$ is to report $\Pi_2 \in \mathcal{R}_{\Pi_1} \cap \mathcal{C}_{\Pi_\Omega}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta$. In any SPE they will do so if that set is non-empty. If there is no such partition, then the second agent will report $\Pi_2 = \Pi_1$, which gets them a score of 0, higher than they would receive for any other report as $S(\Pi_1) > S(\Pi')$ for all $\Pi' \in \mathcal{R}^{>}_{\Pi_1}$.

In anticipation of this behavior, the first agent will not report any $\Pi_1 \in \mathcal{C}_{\Pi_\Omega}$ where $\exists \Pi_2 \in \mathcal{R}_{\Pi_1}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta$. Doing so would result in a score below 0, lower than the minimum score for any partition where no such refinement exists. As reporting any other Π_1 will give the first agent a score of $S(\Pi_1)$, since no refinement can sufficiently change the decision, the first agent will optimize that score by choosing the simplest partition that meets the constraint. Any Π_1 that has been so chosen is definitionally in Π^*.

Given that the first agent will report some $\Pi_1 \in \Pi^*$,and there will not exist any refinements which would change the action chosen to one a distance greater than δ from $D(P_{\Pi_1})$, the second agent will report $\Pi_2 = \Pi_1$. This is an SPE, so one always exists, and if either agent does not follow this behavior, it is not a SPE. □

Theorem 5.3. *In the rebuttal mechanism, an SPE exists, and in any SPE* $\Pi_3 = \Pi_2 = \Pi_1 \in \Pi^{*+}$.

Proof. We prove this with backward induction. Starting with Π_3, the first agent's score is higher by 2 if $d(D(P_{\Pi_3}), D(P_{\Pi_1})) < \delta$. If $(D(P_{\Pi_2}), D(P_{\Pi_1})) < \delta$, then the agent will report $\Pi_3 = \Pi_2$, as this is the simplest partition available in $\mathcal{R}_{\Pi_2} \cap \mathcal{C}_{\Pi_\Omega}$, and will otherwise choose among the simplest partitions that lead to a sufficiently close decision if any exist.

Knowing this and given Π_1, the only way for the second agent to achieve their highest score of $S(\Pi_2) - \frac{S(\Pi_1)+S(\Pi_3)}{2} + 2$ is to report $\Pi_2 \in \mathcal{R}_{\Pi_1} \cap \mathcal{C}_{\Pi_\Omega}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi'})) \geq \delta$ for all $\Pi' \in \mathcal{R}_{\Pi_2} \cap \mathcal{C}_{\Pi_\Omega}$. In any SPE they will do so if that set is non-empty, choosing the simplest among them. If there is no such partition, then the second agent will report $\Pi_2 = \Pi_1$, which will get them a score of 0 once it triggers $\Pi_3 = \Pi_2$, higher than they would receive for any other report as $S(\Pi_1) > S(\Pi')$ for all $\Pi' \in \mathcal{R}^{>}_{\Pi_1}$.

In anticipation of this behavior, the first agent will not report any $\Pi_1 \in \mathcal{C}_{\Pi_\Omega}$ where $\exists \Pi_2 \in \mathcal{R}_{\Pi_1}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi'})) \geq \delta$ for all $\Pi' \in \mathcal{R}_{\Pi_2} \cap \mathcal{C}_{\Pi_\Omega}$. Doing so would result in a score for them of $\frac{S(\Pi_1)+S(\Pi_3)}{2} - 2$ after later reporting $\Pi_3 = \Pi_2$, which is lower than the minimum score of $\frac{S(\Pi_1)+S(\Pi_3)}{2}$ for any report where no such refinement exists. The first agent will then optimize that score by choosing the simplest partition that meets the constraint and plan to set $\Pi_3 = \Pi_2$. Any Π_1 that has been so chosen is definitionally in Π^{*+}.

Given that the first agent will report some $\Pi_1 \in \Pi^{*+}$, and there will not exist any refinements for which all further refinements would change the action chosen to one a distance greater than δ from $D(P_{\Pi_1})$, the second agent will report $\Pi_2 = \Pi_1$, and the first agent will report $\Pi_3 = \Pi_2$. This is an SPE, so one always exists, and if either agent does not follow this behavior, it is not a SPE. □

Theorem 5.4. *For the zero-sum second opinion mechanism under Conditions 1 and 2, an SPE exists, and in any SPE agent* 1 *reports some* $\Pi_1 \in \Pi^*$*, agent* 2 *responds with a prediction,* $\mathcal{D}(P^1_{\Pi_1}, P^2_{\Pi_1}) = a^*_{\Pi_1}$*, and both agents are strictly incentivized to predict honestly for* $a^*_{\Pi_1}$ *and weakly incentivized to predict honestly for all actions.*

Proof. We prove this with backward induction over the game tree. Let $\tilde{a} = D(P^1, P^2)$ denote the action the principal chooses under the agents' reports.

We first show that in any SPE both agents predict honestly on the partition over which they jointly predict. By the analysis of Hudson (2025), the prediction component of agent i's score is maximized in expectation by $p^i_{\tilde{a}} = \mu_{\tilde{a}}$ whenever agent j predicts honestly, since s is strictly proper. For actions $a \neq \tilde{a}$, ω is not observed, so the prediction component does not depend on p^i_a, and by Condition 1 the

simplicity component does not either. The only remaining channel by which a prediction for some action a can affect agent i's score is by changing which action is $\tilde{a}$, and thereby changing the second opinion component. The scaling α_i is constructed so that the expected loss from dishonest reporting on $\tilde{a}$ strictly exceeds the maximum swing of the second opinion component, so this is never a profitable deviation. Therefore both agents are strictly incentivized to predict honestly for $\tilde{a}$ and weakly incentivized to predict honestly for all actions, and when both do so $D(P^1_\Pi, P^2_\Pi) = D(P_\Pi)$ for every partition Π.

After agent 2 reports a strict refinement $\Pi_2 \in \mathcal{R}^{>}_{\Pi_1}$ and prediction $P^2_{\Pi_2}$, agent 1 reports $P^1_{\Pi_2}$, and by the argument above does so honestly in any SPE.

Given Π_1 and $P^1_{\Pi_1}$, agent 2 chooses between reporting only a prediction $P^2_{\Pi_1}$, which is treated as $\Pi_2 = \Pi_1$, and reporting a strict refinement $\Pi_2 \in \mathcal{R}^{>}_{\Pi_1}$ with a prediction $P^2_{\Pi_2}$. Since both agents predict honestly in any SPE, the second opinion component reduces to the original second opinion mechanism applied to the truthful prediction P. The only way for agent 2 to achieve a score above 0 is to report $\Pi_2 \in \mathcal{R}^{>}_{\Pi_1}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta$, in which case the score is $S(\Pi_2) - S(\Pi_1) + 2 > 0$. In any SPE agent 2 will do so if such a partition exists, choosing the simplest. If there is no such partition, agent 2 will report only a prediction, receiving a score of 0, higher than the $S(\Pi_2) - S(\Pi_1) < 0$ they would receive for any strict refinement that does not change the decision.

In anticipation of this behavior, agent 1 will not report any $\Pi_1 \in \mathcal{C}_\Omega$ where $\exists \Pi_2 \in \mathcal{R}^{>}_{\Pi_1}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta$. Doing so would result in a score of $S(\Pi_1) - 2 < 0$, lower than the minimum score of $S(\Pi_1) \geq 0$ for any partition where no such refinement exists. By Condition 2 such partitions are guaranteed to exist, since along any chain of refinements decisions converge, so some sufficiently fine partition lies in $\mathcal{R}^S_\delta$. As reporting any other Π_1 will give agent 1 a score of $S(\Pi_1)$, they will optimize that score by choosing the simplest partition that meets the constraint. Any Π_1 that has been so chosen is definitionally in Π^*.

Given that agent 1 will report some $\Pi_1 \in \Pi^*$, and there will not exist any refinements that change the action chosen to one a distance greater than δ from $D(P_{\Pi_1})$, agent 2 will respond with only a prediction $P^2_{\Pi_1}$, and the principal will choose $D(P^1_{\Pi_1}, P^2_{\Pi_1}) = a^*_{\Pi_1}$. This is an SPE, so one always exists, and if either agent does not follow this behavior, it is not an SPE. □

6. Appendix B: Applications

In this section, we discuss possible applications for the mechanisms we introduce, with a focus on the distance metrics and thresholds used to measure whether the principal changes their mind sufficiently.

6.1. Stochastic Choice

While the principal can only elicit predictions for a finite number of actions, they may be able to choose stochastically over that set, so that their choice is in $\Delta\mathcal{A}$. Distance metrics can then be defined between distributions over actions, such as Euclidian distance or Jensen-Shannon distance.

In practice, even if the decision maker is making stochastic choices, they may be unable to accurately report that distribution. To address this, we can instead use a prediction of their action, elicited from a different expert. This is most plausible in the AI case, where an AI predicting the decision can be consulted frequently and cheaply. Using an AI to predict the decision is also important for training AI experts to simplify, as it makes expected reward continuous and so allows for gradient-based optimization.

6.2. Social Choice

Where we typically think of a single decision making principal, there may be multiple decision makers whose preferences are aggregated into a single decision. Then, distance metrics can incorporate how many decision makers change their mind (and by how much) rather than only applying to the overall aggregated decision. This is helpful for identifying differences in values between decision makers, isolating it for differences in beliefs.

6.3. Histories and Non-Predictions

The outcomes being predicted are often the final states, but the histories that lead to them can also matter to the principal. In some cases, the history is not valued in itself, but provides context on interpreting the final state. There is no issue with allowing predictions over histories, rather than only the final states.

These mechanisms can also be applied to deterministic histories, rather than predictions which give a distribution over histories. This is the special case, where a prediction puts full probability on a single outcome. When this is done, agents are incentivized to simplify the history so as to direct the principal's attention to the relevant events that occur.

6.4. Rating and Ranking

With predictions elicited over actions in $\mathcal{A}$, the principal may wish to rate them or rank them. For example, this is done when training AI systems using reinforcement learning from human feedback (RLHF). Distance metrics can then be applied to changes in ratings, taking the mean or max change across actions. Rankings can be evaluated with metrics like

Kendall tau distance, or Spearman's footrule distance.

6.5. Action Generation

Rather than providing partitions and possibly predictions over a given set of actions, the mechanisms we discuss could be modified so that agents provide a finite set of actions whenever they provide a partition. This would be helpful in cases where experts are aware of actions that the principal would prefer, but that they had not thought of, such as when there is a large space of actions to search through. For this application, the distance metrics work as normal, and would be applied to the principal's choice from a new set of actions.