Formal Incentives for Simplifying Predictions

Rubi Hudson*

January 8, 2026

Abstract

When a principal faces a choice between actions, deciding based on expert predictions of outcomes can still require the principal to expend large amounts of cognitive effort processing them. We investigate mechanisms by which the principal can incentivize an agent to simplify the predictions, while still guiding the principal to their best action. Our focus is on the use of AI agents, which can allow approaches beyond what is possible with humans. We first show that this can be done with a setup using two agents. This is done by rewarding the second agent for providing additional information that sufficiently changes the principal's decision, where the principal controls the bar for sufficiency. 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.

Keywords: predictions, conditional predictions, simplifying predictions

JEL Codes: C53, D82, D86

*University of Toronto

1 Introduction

When making a decision between possible courses of action, a common step is 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. However, this introduces the issue of prediction complexity. While it is possible to incentivize accurate predictions from an expert using a proper scoring rule [Brier, 1950, Good, 1952] regardless of the number of possible outcomes, for a complicated prediction it can be difficult to identify what the relevant outcomes should be. In a setting with ten possible outcome variables that can each take ten possible values, which is far simpler than the real world, predictions over the resulting 10^{10} outcomes are far too large to evaluate. To avoid the extensive process of considering all outcomes, we might want either the expert who provided the prediction or some independent agent to simplify the prediction on our behalf.

To illustrate, consider the case of a patient at the hospital with appendicitis, choosing between undergoing an appendectomy or treating it with antibiotics. Based on the patient's characteristics, a medical professional can predict the likelihood of various outcomes given each course of treatment. However, surgery and antibiotics have dozens of potential side effects, many of which have several degrees of severity. Even if the medical professional could convey their prediction over all outcomes to the patient, the combinatorially large size would be overwhelming. Rather, the patient would prefer that outcomes be clustered together, such as those with side effects of similar type, severity, and/or time frame.

Suppose that a principal hires an agent to simplify a prediction by grouping together outcomes, where simplicity is measured with some function that strictly increases in value when any two outcomes are merged. This function might depend on the actual probabilities involved, so that merging two outcomes that each have high probability affects simplicity differently than merging two outcomes that each have low probability. If the agent is rewarded only for simplicity, the optimal report is a single group containing all outcomes. The principal would then be choosing with no information, which could result in taking a

suboptimal or even catastrophic action.

What the principal would prefer is to simplify the prediction only insofar as it allows them to still choose their best action, the one which they would take if they spent the effort to process the full information. However, if the outcome space is large enough that simplification is useful, it will usually be too large to contract over, meaning the principal cannot align the agent by rewarding them proportionally to the value of the outcome realized.

1.1 Our Contribution

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 simplified grouping of outcomes, the principal allows the other agent to provide a 'second opinion'. If the second agent can further split outcomes in a way that changes the principal's choice of action, the second agent is rewarded and the first is punished. This makes it so that the first agent is incentivized to simplify the original prediction only insofar as it does not change the principal's decision. Following up on this mechanism, we provide additional results showing that we can elicit further simplification while still choosing the optimal action with a mechanism that allows the first agent to reply, and that either mechanism can be modified to be made symmetric and/or simultaneous.

We then examine the case where there does not exist an expert prediction to be simplified, but rather one must be elicited in simplified form. In many applications, the set of outcomes to be predicted over is too large for an expert to provide a prediction on, notably the cases where outcome variables take on continuous values. While the principal can decide on the set of outcomes before eliciting predictions, this may cause them to miss crucial information that experts are aware of. Instead, we would like the expert, who knows all the relevant information, to come up with the ideal simplification of outcomes and report the probabilities over that set. We show how the previous mechanism can be modified so that this outcome is incentivized.

1.2 Literature Review

The literature on eliciting predictions over a set of outcomes is extensive, starting with [Brier, 1950] and [Good, 1952]. [Gneiting and Raftery, 2007] establishes properties that apply to all (strictly) proper scoring rules, which are defined as those that (strictly) incentivize reporting actual beliefs.

Eliciting multiple conditional predictions from a single expert for use in making decisions was investigated by [Othman and Sandholm, 2010], who showed that because predictions for untaken actions cannot be evaluated, it is impossible to use predictions from a single agent to deterministically identify and take the best action available. Follow-up work [Chen et al., 2011, Oesterheld and Conitzer, 2020] found partial workarounds in the single expert case, and [Hudson, 2025] showed that multiple agents in a zero-sum competition can be incentivized in a way that circumvents the impossibility result. While these papers use a similar model to our proposed research direction, they focus on eliciting accurate predictions rather than simplifying them. To our knowledge, we are the first to explore the direction of simplifying prediction.

The rational inattention literature, starting with Sims [2003], models individuals as having a limited capacity to process information, and needing to allocate their attention efficiently. Gabaix [2014] connected this model to broader theories of bounded rationality, while Matějka and McKay [2015] applied it to a discrete choice model. The rational inattention literature is descriptive, aiming to model how humans typically act, rather than prescriptive like our work. However, it provides powerful motivation, showing that there is a need for predictions to be simplified before they can be used in decision making. The novelty of our approach is in using an outside agent for this simplification.

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]. Our proposed work differs in that arguments are restricted to

be true simplifications of a set of predictions, and that the reward for changing the principal's mind only applies off the equilibrium path.

Early work by Blackwell [1951] established a foundation for comparing information structures, showing that one is more informative than another if and only if it leads to better decisions for any decision problem and preferences. In contrast, our work is aimed at eliciting the information structure that leads to the best decision for a specific decision problem and preferences, and penalizes the more complex information structures that help with others.

The closest paper to our work is Lipnowski et al. [2020], which modeled the case where a principal without attention cost summarizes 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 over outcomes, while our interest is the case where the decision maker can set the utility function of the summarizer but cannot contract over their full utility function.

2 Model

Let $\mathcal{A}$ be a finite set of actions, and let Ω be an exhaustive and mutually exclusive 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. That is, for each $\pi_1 \in \Pi_1$, there exists $\pi_2 \in \Pi_2$ such that $\pi_1 \subseteq \pi_2$. A coarsening is *strict* if $|\Pi_2| < |\Pi_1|$, and a refinement is *strict* if $|\Pi_1| < |\Pi_2|$. Let $\mathcal{C}_\Pi$ denote all coarsenings of Π and $\mathcal{C}_\Pi^>$ the strict ones. Similarly, let $\mathcal{R}_\Pi$ denote all refinements of Π and $\mathcal{R}_\Pi^>$ the strict ones. We abuse notation slightly and 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 are interested in a reward that can be costlessly provided by the principal. These agent preferences are standard for prediction scoring rules. 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 $S_{\text{ent}}(\Pi; P) = \sum_{a \in \mathcal{A}} \sum_{\pi \in \Pi} p^{\Pi}_{a,\pi} \log p^{\Pi}_{a,\pi}$ (log base 2 is used for Shannon entropy). 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. By default, we think of this as the discrete metric, such that $d(x, y) = 0$ if $x = y$ and 0 otherwise, but some applications suggest a natural metric for distance.

Finally, we have a decision-making principal with complete and transitive preferences $\succsim$ over $\bigcup_{\Pi \in \mathcal{C}_\Omega} \Delta(\Pi)$, i.e., over all distributions induced by possible partitions of Ω. For simplicity, we assume a tie-breaking procedure (including across actions yielding the same distribution), so preferences are strict. Let $D : \bigcup_{\Pi \in \mathcal{C}_\Omega} \Delta(\Pi)^{|\mathcal{A}|} \to \mathcal{A}$ map a set of conditional predictions to the action corresponding to the most preferred distribution. Agents are assumed to know the principal's preferences, since the principal will not want to conceal them.

3 Results

3.1 Simplifying Predictions

We first analyze the case where an initial expert prediction P_Ω on a partition Π_Ω is given. The principal would like to be provided with the simplest information set such that no additional information will change their choice of action to one with a distance greater than δ. Under the discrete distance metric, this is accomplished with the simplest partition Π such that $d(D(P_{\Pi_R}), D(P_\Pi)) < \delta$ for all $\Pi_R \in \mathcal{R}_\Pi$. We will call the set of all such partitions Π^*, noting that there may be multiple tied for the simplest.

It is straightforward to argue that no scoring rule can always incentivize an expert to provide a partition in Π^*. For some predictions, the optimal partition would be the maximal coarsening, which groups all outcomes together, if the principal's decision upon seeing that happens to be correct. However, upon receiving such a partition from an expert, the principal has no way to evaluate whether it is in Π^*. Any scoring rule that incentivizes reporting it therefore does so independently of whether it is in Π^*, and so cannot incentivize only reporting it in that case.

Fortunately, we can show that what we cannot incentivize from a single expert, we can incentivize from a pair of experts.

3.1.1 Sequential Elicitation

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, $\Pi_2 \in \mathcal{R}_{\Pi_1} \cap \mathcal{C}_{\Pi_\Omega}$. If $d(D(P_{\Pi_1}), D(P_{\Pi_2})) < \delta$, rewards are $(S(\Pi_1), S(\Pi_2) - S(\Pi_1))$, and if $d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta$ then rewards are $(S(\Pi_1) - 2, S(\Pi_2) - S(\Pi_1) + 2)$. As the principal's actual decision does not affect the incentives of the agents, we can leave their implemented choice when $D(P_{\Pi_1}) \neq D(P_{\Pi_2})$ undetermined. It is not necessary to subtract $S(\Pi_1)$ from the score of agent 2 here, or change scores by 2 instead of 1 when the decision shifts sufficiently,

but it becomes helpful in later extensions.

When the second partition does not shift the principal's decision beyond the threshold amount, the first agent's reward is equal to the simplicity of the partition they provided, while the second agent's is the negative value of how much their partition decreases the simplicity score from Π_1. If the second agent does not shift the decision, their strictly highest scoring report in $\mathcal{R}_{\Pi_1}$ is $\Pi_2 = \Pi_1$, which avoids the need to evaluate a second partition. When the second partition does sufficiently shift the principal's decision, the first agent's score drops by 2, making it lower than any score they could receive if the decision does not shift by at least δ, while the second agent's score is higher.

We can show that this mechanism incentivizes both agents to report a partition in Π^*.

Theorem 1. *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 $S(\Pi_1)$, 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. □

In addition to incentivizing the simplest partition such that the choice of action will be within δ of the choice induced by any refinements, we may also be interested in the looser condition of the choice of action being within δ of $D(P_{\Pi_\Omega})$. In this case, it would be acceptable if refinement would lead to a very different decision, so long as further refinement would bring it back. This is done with the simplest partition Π such that $d(D(P_{\Pi_\Omega}), D(P_\Pi)) < \delta$. The set of all such partitions will be called Π^{*+}, noting again that there may be multiple tied for the simplest.

To elicit a partition in Π^{*+}, consider the following setup, which we call the *debate 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 $\Pi_3 \in \mathcal{R}_{\Pi_2} \cap \mathcal{C}_{\Pi_\Omega}$. If $d(D(P_{\Pi_3}), D(P_{\Pi_1})) < \delta$, rewards are $(\frac{S(\Pi_1)+S(\Pi_3)}{2}, S(\Pi_2) - \frac{S(\Pi_1)+S(\Pi_3)}{2})$ and when $d(D(P_{\Pi_3}), D(P_{\Pi_1})) \geq \delta$, rewards are $(\frac{S(\Pi_1)+S(\Pi_3)}{2} - 2, S(\Pi_2) - \frac{S(\Pi_1)+S(\Pi_3)}{2} + 2)$.

When the third partition does not shift the principal's decision from the first beyond the threshold amount, the first agent's reward is equal to the average simplicity of their partitions. The second agent's is simply the negative value of how much the decreased the simplicity score from Π_1. When the third partition does shift the principal's decision sufficiently far from the first, the first agent's score is lowered by 2, making it lower than any score they could receive if the decision does not shift, and the second agent's score is higher. We can show that this mechanism incentivizes both agents to report a partition in Π^{*+}.

Theorem 2. *In the debate 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. □

3.1.2 Simultaneous Elicitation

The second opinion mechanism and the debate mechanism are presented above as asymmetric and sequential. This is based on the motivating case of having one agent review the work of the other, which for some applications would be easier to implement. However, we can also elicit partitions in Π^* and Π^{*+} simultaneously in equilibrium and symmetrically, when doing so is more practical.

For the Π^* case, consider the mechanism where agents report partitions Π_1 and Π_2 simultaneously. If $S(\Pi_1) = S(\Pi_2)$, then scores are $(0, 0)$. Otherwise, if $S(\Pi_i) > S(\Pi_j)$, the second opinion mechanism is run starting with agent i reporting Π_i, and letting agent j respond with a partition. Final scores are the scores from that second opinion mechanism, with

$S(\Pi_j)$ subtracted from agent i's final score and added to agent j's final score. We call this the *simultaneous second opinion mechanism.*

Theorem 3. *In the simultaneous second opinion mechanism, an SPE exists, and in any SPE* $\Pi_1, \Pi_2 \in \Pi^*$ *so that the second opinion mechanism is not triggered.*

Proofs for this and later results are provided in Appendix A.

An analogous result holds for Π^{*+}, running the debate mechanism on the simpler partition instead of the second opinion mechanism, with an analogous proof. We call that the *simultaneous debate mechanism.*

These mechanisms also have the useful property that they are zero-sum, making it impossible for the agents to collaborate with each other. On the other hand, a downside of simultaneous elicitation is that both agents can report different partitions in Π^*, even in equilibrium, which would require the principal to evaluate a second distinct partition.

3.1.3 Memoryless Elicitation

An advantage of using an AI to simplify predictions 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. To do this, we set $\Pi_0 = \Pi_\alpha$, the maximal coarsening, such that is consists of a single outcome occurring with probability 1. Agent i is shown Π_{i-1} and asked to produce a coarsening Π_i. If $\Pi_{i-1} = \Pi_i$, the process ends, otherwise it proceeds to agent $i+1$.

To elicit a partition in Π^*, we can assign agent i a score of

$$S(\Pi_i) - S(\Pi_{i-1}) + \mathbb{I}(d(D(P_{\Pi_{i-1}}), D(P_{\Pi_i})) \geq \delta)[1 + S(\Pi_{i-1}) - \mathbb{I}(d(D(P_{\Pi_i}), D(P_{\Pi_{i+1}})) \geq \delta)[(1 + S(\Pi_{i+1}))]]$$

If we force the process to stop after agent 2, then scores for agents 1 and 2 will be

$$((S(\Pi_1) - (1 + S(\Pi_2)) * \nVdash(d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta),$$

$$S(\Pi_2) - S(\Pi_1) + (1 + S(\Pi_2)) * \nVdash(d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta)))$$

which creates the same incentives as the second opinion mechanism. As such, we call this the *memoryless second opinion mechanism.*

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.

Theorem 4. *In the memoryless second opinion mechanism, an SPE exists, and in any SPE* $\Pi_i, \in \Pi^*$ *for all* i*.*

To elicit a partition in Π^{*+}, we can assign agent i a score of

$$S(\Pi_i) - S(\Pi_{i-1}) + \mathbb{I}(d(D(P_{\Pi_{i-1}}), D(P_{\Pi_i})) \geq \delta)[1 - \mathbb{I}(d(D(P_{\Pi_i}), D(P_{\Pi_{i+1}})) \geq \delta)]$$

If we force the process to stop after agent 3, and combine the scores for agent 1 and agent 3 (if they are reached), then scores for agents 1 and 2 will be

$$((S(\Pi_1) + S(\Pi_3) - S(\Pi_2) - \nVdash(d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta)[1 - \nVdash(d(D(P_{\Pi_2}), D(P_3)) \geq \delta)],$$

$$S(\Pi_2) - S(\Pi_1) + \nVdash(d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta)[1 - \nVdash(d(D(P_{\Pi_2}), D(P_3)) \geq \delta)]), D(P_{\Pi_2})) \geq \delta)))$$

This creates the almost the same incentives as the debate mechanism, except that agent 3 only tries to change the decision away from $D(P_{\Pi_2})$, rather than toward $D(P_{\Pi_1})$), which can be easily changed if desired . As such, we call this the *memoryless debate mechanism.*

The difference between these mechanisms is that the memoryless second opinion mechanism rewards an additional $S(\Pi_{i-1})$ for changing the decision sufficiently from the previous agent, and subtracts an additional $S(\Pi_{i+1})$ if the following agent changes the decision sufficiently. The effect of this is that if both occur, the agent receives a score of $S(\Pi_i) - S(\Pi_{i+1})$, which is greater than the 0 they would receive for reporting $\Pi_i = \Pi_{i-1}$. Under the memoryless debate mechanism, the score for an agent who changes the decision and then has it changed by the following agent is instead $S(\Pi_i) - S(\Pi_{i-1}) < 0$. So, agents will take into account 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 debate mechanism.

3.2 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 $\mathcal{D}(P^i, P^j)$ and observing outcome ω the score for agent $i \neq j$ is given by $\mathcal{S}_i(P^i, P^j, \omega) = s(p^i_{\mathcal{D}(P^i,P^j)}, \omega) - s(p^j_{\mathcal{D}(P^i,P^j)}, \omega)$, 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 across both agents, even if they disagree for that action. While both agents predict simultaneously in that setup, it does not affect their incentives to have them do so sequentially instead.

The second opinion or debate mechanism 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^{*+}. However, this will only work for certain measures of simplicity. The simplicity measure cannot change based on the predicted distributions for untaken actions, as this would violate the weak incentive for honesty on untaken actions. Notably, the negative of the number of elements in a partition is a simplicity measure that does not depend on predictions for untaken actions.

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

With continuous Ω, the principal may have preferences such that for any partition Π, there exists a partition $\Pi' \in \mathcal{R}_\Pi$ where $d(D(P_\Pi), D(P_{\Pi'})) \geq \delta$. In that case, providing more information never causes the principal's decision to converge to a particular action, making the question of what information is relevant meaningless. To rule this out, we add the following condition:

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 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 provides 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 second opinion component of the score 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 $\alpha_i(P^i, P^j, \mathcal{D}(P^i, P^j))(s(P^i, \omega) - s(P^j, \omega))$, where s is a strictly proper scoring rule and P^i is the prediction that is made by agent i over the same partition as agent j. α_i is a function for scaling the prediction score, so that the incentive for honesty dominates the incentives of the second opinion mechanism. If predictions are made on Π_1, then it takes on value 1, and if predictions are made on Π_2 and $p^i_{\mathcal{D}(P^i,P^j)} = p^j_{\mathcal{D}(P^i,P^j)}$, then it takes on value 0. Otherwise,

$$\alpha_i(P^i, P^j, \mathcal{D}(P^i, P^j)) = \frac{2}{E_{\omega \sim p^j_{\mathcal{D}(P^i,P^j)}}[s(p^i_{\mathcal{D}(P^i,P^j)}, \omega) - s(p^j_{\mathcal{D}(P^i,P^j)}, \omega)]} + 1$$

This makes it so that the expected loss from making a dishonest prediction while the other agent predicts honestly is greater than 1, which offsets any gain they could make by increasing the apparent distance between actions.

When the above conditions hold, the zero-sum second opinion mechanism elicits quasi-strictly proper prediction over a partition in Π^*.

Theorem 5. *For the zero-sum second opinion mechanism under Conditions 1 and 2, a sub-game perfect equilibrium (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.*

An analogous result again holds for Π^{*+}, running the debate mechanism on the simpler partition instead of the second opinion mechanism, with an analogous proof. We call that the *zero-sum debate mechanism.* Much like the initial second-opinion and debate mechanisms, the zero-sum versions can also be modified to be symmetric and simultaneous.

4 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.

4.1 Stochastic Choice

The principal can only elicit predictions for a finite number of actions, but 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 choices stochastically, 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 changes in decisions continuous, which allows for gradient-based optimization.

4.2 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 particularly helpful for identifying "cruxes" between multiple decision makers, highlighting what the important considerations are for each of them and allowing them to then discuss the exact issues that drive their differences. This can also be used to specifically identify differences in values between decision makers, isolating them from differences in beliefs about the likelihoods of various outcomes.

4.3 Histories and Non-Predictions

The outcomes being predicted are often the final states of the world, 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 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 effectively the prediction puts full probability on a single history. When this is done, the benefit is that agents are incentivized to simplify the history to direct the principal's attention to the relevant events that occur.

4.4 Rating and Ranking

With predictions elicited over actions in $\mathcal{A}$, the principal may wish to rate them or rank them, and not merely choose the best. For example, rating actions provides more information when done for reinforcement learning purposes, training an AI to take act in support of the principal's preferences. 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.

4.5 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.

5 Discussion and Future Work

We have presented a number of mechanisms for using multiple experts to simplify predictions for a decision maker. The mechanisms are generally not unique, and so it may be possible to streamline them further, or modify them so that they display additional properties desirable for some application. Our work outlines proofs of concept, showing that simplification is possible and providing a concrete starting point for experimental testing.

Empirical evaluation to build on our theoretical results is currently ongoing. We divide our experiments into eliciting simplified predictions, and using simplified predictions. Eliciting simplified predictions is be tested straightforwardly, by having one Large Language Model (LLM) decide between actions based on predictions of their outcomes that have been made and simplified by another LLM. Preliminary results suggest that the using the second-opinion mechanism and its variants as reward functions train LLMs to simplify predictions, but only insofar as it continues to provide the decision maker with necessary information.

For experiments on the use of simplified predictions, we are particularly enthusiastic regarding applications related to the predictions used in training LLMs. In an actor-critic setup, an actor head on a neural network is trained to take actions that the critic head predict will lead to high reward. However, the critic's prediction is a single value, the average expected reward. With our methods we can break the prediction down into uncertainty over outcomes, and uncertainty over reward given an outcome. This would allow for either training or constraining an LLM agent to take more conservative actions when it was significant uncertainty as to what the outcome will be, as well as targeted value learning at the outcomes where it is uncertain regarding reward.

Our work defining formal incentives for simplifying predictions bypasses an enormous restriction on the use of predictions for decision making. From humans making better decisions with the use of AI to provide information, to AI making better decisions based on being more aligned to humans, this work has the potential for improving the crucial decisions that will get made as our world becomes more integrated with powerful AI systems.

6 Acknowledgment

We thank the Cosmos x Fire grants program for supporting this work.

References

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

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

Yiling Chen, Ian Kash, Mike Ruberry, and Victor Shnayder. Decision markets with good incentives. In *International Workshop on Internet and Network Economics*, pages 72–83. Springer, 2011.

Xavier Gabaix. A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129(4):1661–1710, 09 2014. ISSN 0033-5533. doi: 10.1093/qje/qju024. URL https://doi.org/10.1093/qje/qju024.

Tilmann Gneiting and Adrian 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.

Katja Grace. 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.

Rubi Hudson. 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.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

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

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *The American Economic Review*, 105(1):272–298, 2015. URL http://www.jstor.org/stable/43497060.

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

Abraham Othman and Tuomas 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. Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS), 2010.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003. ISSN 0304-3932. doi: https://doi.org/10.1016/S0304-3932(03)00029-1. URL https://www.sciencedirect.com/science/article/pii/S0304393203000291. Swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information.

Appendix A: Proofs

Theorem 3. *In the simultaneous second opinion mechanism, an SPE exists, and in any SPE* $\Pi_1, \Pi_2 \in \Pi^*$ *so that the second opinion mechanism is not triggered.*

Proof. Without loss of generality, assume $S(\Pi_1) \geq S(\Pi_2)$ if $S(\Pi_i) > S(\Pi_j)$. When the second opinion game is played out starting with Π_1, equilibrium scores are $(S(\Pi_1), 0)$ if there does not exist $\Pi_2' \in \mathcal{R}_{\Pi_1} \cap \mathcal{C}_{\Pi_\Omega}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi_2'})) \geq \delta$ and $(S(\Pi_1) - 2, S(\Pi_2') - S(\Pi_1) + 2)$ otherwise. As such, if $S(\Pi_1) > S(\Pi_2)$ and $\Pi_1 \in \Pi^*$, total scores will be $(S(\Pi_1) - S(\Pi_2), S(\Pi_2) - S(\Pi_1))$, which cannot be an equilibrium as the second agent could increase their score to 0 by instead reporting $\Pi_2 = \Pi_1$.

If $S(\Pi_1) > S(\Pi_2)$ and $\Pi_1 \notin \Pi^*$, then either there exists $\Pi_2' \in \mathcal{R}_{\Pi_1} \cap \mathcal{C}_{\Pi_\Omega}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi_2'})) \geq \delta$ or not. If there exists such a partition, then final scores will be $(S(\Pi_1) - 2 - S(\Pi_2), S(\Pi_2') + 2 + S(\Pi_2))$, which cannot be an equilibrium as the first agent could increase their score to 0 by reporting $\Pi_2 = \Pi_1$. If there does not exist such a partition, then final scores will be $(S(\Pi_1) - S(\Pi_2), S(\Pi_2) - S(\Pi_1))$, where either $\Pi_2 \in \Pi^*$ and agent 1 could increase their score to 0 by reporting $\Pi_2 = \Pi_1$ or $\Pi_2 \notin \Pi^*$ and agent 2 could increase their score to by reporting $\Pi_2 \in \Pi^*$. Either way, these cannot equilibria.

If $S(\Pi_1) = S(\Pi_2)$ and $\Pi_1 \notin \Pi^*$, then if $S(\Pi_2) < S(\Pi')$ for $\Pi' \in \Pi^*$, agent 2 can increase their score by reporting $\Pi_2 = \Pi'$. If $S(\Pi_2) \geq S(\Pi')$ for $\Pi' \in \Pi^*$, agent 2 can increase their score by reporting $\Pi_2 \in \mathcal{R}^{>}_{\Pi_1}$ triggering the second opinion mechanism starting with Π_1. Therefore, neither of these can be equilibria.

Now we show that $\Pi_1, \Pi_2 \in \Pi^*$ are equilibria. For all Π_i' such that $S(\Pi_i') > S(\Pi_i)$, by the definition of Π^* there exists $\Pi_j' \in \mathcal{R}_{\Pi_i'} \cap \mathcal{C}_{\Pi_\Omega}$ such that $d(D(P_{\Pi_i}), D(P_{\Pi_j'})) \geq \delta$, so switching would trigger the second opinion mechanism and result in a lower score. For all Π_i' such that $S(\Pi_i') < S(\Pi_i)$, switching results in a lower final score of $S(\Pi_i') - S(\Pi_j)$. And for all Π_i' such that $S(\Pi_i') = S(\Pi_i)$ but $\Pi_i' \notin \Pi^*$, switching results in the same final score of 0. As such, there are no profitable deviations for either agent, making $\Pi_1, \Pi_2 \in \Pi^*$ equilibria, and since

$S(\Pi_1) = S(\Pi_2)$, the second opinion mechanism is not run. □

Theorem 4. *In the memoryless second opinion mechanism, an SPE exists, and in any SPE* $\Pi_i, \in \Pi^*$ *for all* i.

Theorem 5. *For the zero-sum second opinion mechanism under Conditions 1 and 2, a subgame perfect equilibrium (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. If agent 1 reports $\Pi_1 \in \Pi^*$ with $P^1_{\Pi_1} = \mu_{\Pi_1}$, and $P^1_{\Pi_2} = \mu_{\Pi_2}$ if necessary, then they are guaranteed a score of at least $S(\Pi_1)$. If agent 2 reports $P^2_{\Pi_1} = \mu_{\Pi_1}$, they are guaranteed a score of at least 0. So, in any equilibrium, both agents receive scores at least that high.

If agent 1 reports Π_1 such that there exists some $\Pi_2 \in \mathcal{R}_{\Pi_1}$ with $d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta$, and agent 2 responds with such a partition and $P^2_{\Pi_2} = \mu_{\Pi_2}$, then the final score of agent 1 is capped at $S(\Pi_1) - 2$. This is lower than if they had reported $\Pi_1 \in \Pi^*$ with $P^1_{\Pi_1} = \mu_{\Pi_1}$, so cannot be an equilibrium. Agent 2 cannot achieve a higher score by responding with a different partition and prediction, so if they do not it must be because they respond with only a prediction $P^2_{\Pi_1}$, and their expected score from doing so is greater than $2 - S(\Pi_1) + S(\Pi_2)$. Then, agent 1's score will be $S(\Pi_1) - [1 - S(\Pi_1) + S(\Pi_2)] = 2S(\Pi_1) - S(\Pi_2) - 2 < 0 < S(\Pi'_1)$ for any Π'_1, so that also cannot be an equilibrium.

If agent 1 reports Π_1 such that $S(\Pi_1) < S(\Pi')$ for some $\Pi' \in \Pi^*$, and there does not exist some $\Pi_2 \in \mathcal{R}_{\Pi_1}$ such that $d(D(P_{\Pi_1}), D(P_{\Pi_2})) \geq \delta$, then the final score of agent 1 is capped at $S(\Pi_1) < S(\Pi'_1)$ for any $\Pi'_1 \in \Pi^*$. This is because if it were higher, then agent 2 must be receiving a score below 0, which cannot occur in equilibrium. As such, this cannot be an equilibrium either, and so it must be that in any equilibrium, $\Pi_1 \in \Pi^*$. If $\Pi_1 \in \Pi^*$ but $D(P^1_{\Pi_1}) \neq a^*_{\Pi_1}$, then agent 2 can maximize their score by reporting $P^2_{\Pi_1} = \mu_{\Pi_1}$, which as shown in Hudson [2025] will lower agent 1's total score below $S(\Pi')$ for some $\Pi' \in \Pi^*$, so in any equilibrium we also have that $D(P^1_{\Pi_1}) = a^*_{\Pi_1}$.

When agent 2 responds with a partition $\Pi_2 \in \mathcal{R}^{>}_{\Pi_1}$ and prediction $P^2_{\Pi_2} = \mu_{\Pi_2}$, agent 1's expected score is maximized with a prediction $P^1_{\Pi_2}$ such that $p^1_{\mathcal{D}(P^1_{\Pi_2}, P^2_{\Pi_2}),\Pi_2} = p^2_{\mathcal{D}(P^1_{\Pi_2}, P^2_{\Pi_2}),\Pi_2}$. For any such prediction, agent 2 will receive an expected score of $S(\Pi_2) - S(\Pi_1) < 0$, and so that cannot be an equilibrium. If $P^2_{\Pi_2} \neq \mu_{\Pi_2}$, then if there exists $P^1_{\Pi_2}$ such that $\mathcal{D}(P^1_{\Pi_2}, P^2_{\Pi_2}) = a^*_{\Pi_1}$ and $E_\mu[\alpha_i(P^1_{\Pi_2}, P^2_{\Pi_2}, \mathcal{D}(P^1_{\Pi_2}, P^2_{\Pi_2}))s(P^1_{\Pi_2}, \omega) - s(P^2_{\Pi_2}, \omega)] > 0$ or $\mathcal{D}(P^1_{\Pi_2}, P^2_{\Pi_2}) \neq a^*_{\Pi_1}$ but $E_\mu[\alpha_i(P^1_{\Pi_2}, P^2_{\Pi_2}, \mathcal{D}(P^1_{\Pi_2}, P^2_{\Pi_2}))s(P^1_{\Pi_2}, \omega) - s(P^2_{\Pi_2}, \omega)] > 2$, agent 1 will take such an action. If they do, agent 2 will receive an expected score lower than $S(\Pi_2) - S(\Pi_1) < 0$, so this cannot be an equilibrium. As such, in any equilibrium, agent 2 does not respond with a partition.

If agent 2 responds with a prediction, then as shown in Hudson [2025], we will have that in any equilibrium $\mathcal{D}(P^1_{\Pi_1}, P^2_{\Pi_1}) = a^*_{\Pi_1}$ and $p^1_{\mathcal{D}(P^1_{\Pi_1}, P^2_{\Pi_1}),\Pi_1} = p^2_{\mathcal{D}(P^1_{\Pi_1}, P^2_{\Pi_1}),\Pi_1} = \mu_{\Pi_1}$. The above shows that this is an equilibrium, as any deviation induces a subgame resulting in a lower score for the deviator. □