Extrapolating Volition with Recursive Information Markets

Abhimanyu Pallavi Sudhir *
Department of Computer Science
University of Warwick
Coventry – CV4 7ES, UK.
abhimanyu.pallavi-sudhir@warwick.ac.uk

Long Tran-Thanh
Department of Computer Science
University of Warwick
Coventry – CV4 7ES, UK.
long.tran-thanh@warwick.ac.uk

Abstract

One of the impediments to the efficiency of information markets is the inherent information asymmetry present in them, exacerbated by the "buyer's inspection paradox" (the buyer cannot mitigate the asymmetry by "inspecting" the information, because in doing so the buyer obtains the information without paying for it). Previous work has suggested that using Large Language Model (LLM) buyers to inspect and purchase information could overcome this information asymmetry, as an LLM buyer can simply "forget" the information it inspects. In this work, we analyze this mechanism formally through a "value-of-information" paradigm, i.e. whether it incentivizes information to be priced and provided in accordance with its "true value". We focus in particular on our new *recursive* version of the mechanism, which we believe has a range of applications including in AI alignment research, where it is related to Extrapolated Volition and Scalable Oversight.

1 Introduction

A key challenge in economics is the development of mechanisms for efficiently pricing and supplying information [Stigler, 1961]. Simple, naive information markets suffer from a number of flaws: *the low cost of duplication* [Samuelson and Nordhaus, 2009], *the transaction costs of tenders* [Thomassen et al., 2016], the transaction cost of *learning* new information, and *information asymmetry* and *the buyer's inspection paradox* [Arrow, 1972, Van Alstyne, 1999].

The buyer's inspection paradox refers to the fact that in information markets, a buyer cannot simply inspect some information and decide whether to buy it, because the moment she does so, she has already obtained it and no longer has an incentive to correctly value it. This creates a tradeoff between information asymmetry and positive externalities of information, depending on how much of the information the seller is willing to reveal freely. This flaw is relevant not only to information markets, but even to regular goods markets without perfect information, as famously described in the "Market for Lemons" [Akerlof, 1978]. Barzel [1985] goes further to claim that information market inefficiency is *the* fundamental cause of all market failure.

Recently, Weiss et al. [2024] proposed a novel mechanism to mitigate the buyer's inspection paradox, via the use of Large Language Models: instead of a human buyer directly inspecting the information for purchase, he deploys an LLM agent to inspect the information on his behalf. The LLM agent makes the purchase decision (or recommends a purchase decision without further explanation) and erases the inspected information from its context window.

In this work, we build on their framework, developing a formal mathematical theory for information markets with inspection. In particular, we observe that their framework still does not eliminate

*Part of the work was done while at PIBBSS with mentorship from Deger Turan.

Preprint. Under review.

information asymmetry, because the LLM buyer inspecting information *still* faces information asymmetry.

Our contributions are as follows:

1. **Recursive inspection protocol.** Extending the work of Weiss et al. [2024] to account for the persistence of information asymmetry in the subsidiary information market, we introduce the *Recursive Information Protocol*. We discover that the straightforward method of "simply applying Weiss et al. [2024] to itself" is too simplistic, and instead provide a more robust protocol, formulated as an imperfect-recall game.

2. **Mathematical formalism.** We give a fully formal mathematical description for a Bayesian agent participating in the Recursive Information Protocol, as well as a proof of an optimality result Theorem 2.3 demonstrating that the "Recursive Inspection Protocol is ex-ante superior to any other admissible protocol" (where an admissible protocol is one that respects the property rights of the information sellers).

3. **Practical implementation.** We provide an implementation of an information market server implementing the Recursive Inspection Protocol, detailed in Section 3, which can directly be applied to various practical applications for information markets such as question-and-answer sites, product inspections and online fact-checking.

1.1 Related work

Information economics. The foundational theory of information economics is the *value-of information* framework [Kuhn et al., 1953, Howard, 1966], which treats information as an instrumental good [Van Alstyne, 1999]. The buyer's inspection paradox was introduced by Arrow [1972] and named by Van Alstyne [1999]. Hanson [2011a,b] discussed the inefficiency of information markets in the context of intellectual property (IP) law, commenting: "*just as farmers developed barbed-wire, someday I expect IP advocates will develop better forms of intellectual property*".

Mechanism design for information markets. Conitzer [2009] introduced a mechanism for rewarding information-providing agents based on their influence on prediction market prices.

Scalable oversight The fundamental limitation of RLHF that it relies on a human's ability to judge a (potentially superhuman) AI's outputs has long been recognized [Christiano et al., 2018], and is known as the *scalable oversight* problem in the AI alignment literature [Hubinger, 2020, Bowman et al., 2022]. One well-known scalable oversight proposal is Debate [Irving et al., 2018]; our proposal to augment RLHF with information markets may be seen as yet another such proposal.

Mechanism design with LLMs. Market design for LLM participants has opened up many new frontiers previously not possible with humans alone. Apart from the information bazaar [Weiss et al., 2024], this includes e.g. token auctions for online advertising [Dütting et al., 2024], economic simulations with AI agents [Li et al., 2024], and forecasting with LLMs [Schoenegger et al., 2024, Halawi et al., 2024, Paleka et al., 2024, Buterin, 2024].

Zero-knowledge proofs. Another way for a seller to prove the value of his information without revealing it is via a *zero-knowledge proof* [Goldreich, 2001] – in formal settings, this is available if the information is a solution to a PSPACE problem [Impagliazzo and Yung, 1987, Ben-Or et al., 1990]; extending this to informal settings is an active area of work [Hammond and Adam-Day, 2024].

Miscellaneous Recent works like Manelli and Vincent [2006], Babaioff et al. [2012], Bergemann et al. [2018] have studied "optimal mechanisms for selling information", but from the point-of-view of a seller maximizing his revenue – we, on the other hand, are interested in improving the efficiency of the information market itself. Information market mechanisms have also been designed for *data markets* in machine learning, e.g. Fallah et al. [2024], Chen et al. [2022], Ghorbani and Zou [2019].

2 Bayesian setting

We are concerned with modeling an agent α's "value for information" in an expected utility maximization framework. We assume a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ (with $\mathbf{P}$ a common prior for all agents ever discussed) and that the only source of utility is some decision problem specified by a set of choices $\mathcal{X}$ for α and payoffs given by a measurable utility function $U : \Omega \times \mathcal{X} \to \mathbb{R}$. An "information good" is a tuple of a random variable I, its realization or "true value"[2] $I(\omega) = i$ and a price: $\mathbf{I} = \langle I, i, p \rangle$. All of these contents of an information good may be hidden from α. We want to study how α values informational goods.

With no further information, α's choice would just be $\arg\max_{x \in \mathcal{X}} \mathbf{E}[U(x)]$. With information $\langle I, i, p \rangle$, the agent would instead choose $U(\arg\max \mathbf{E}[U(x) \mid I = i])$. Thus we can say the utility of that information is:

$$U(\mathbf{I}) = U(\arg\max \mathbf{E}[U(x) \mid I = i]) - U(\arg\max \mathbf{E}[U(x)]) - p \tag{1}$$

If α has to decide whether to buy an information good (or decide which one to buy out of a list of offers), it has to take the expectation of $U(\mathbf{I})$. There are several ways to do this. There is the *ex-post* value of information $\mathbf{E}[U(\mathbf{I})|I = i]$, which is α's estimate for $\mathbf{I} = \langle I, i, p \rangle$ after viewing it:

$$\mathbf{E}[U(\mathbf{I})|I = i] = \max_{x \in \mathcal{X}} \mathbf{E}[U(x) \mid I = i] - \mathbf{E}\left[U\left(\arg\max_{x \in \mathcal{X}} \mathbf{E}[U(x)]\right) \mid I = i\right] - p \tag{2}$$

The *ex-ante* value (before seeing the information) may be taken as the expectation over Equation (2): $\mathbf{E}_{i \sim I}[\mathbf{E}[U(\mathbf{I})|I = i]]$ (which would be the *value of an experiment* [Lindley, 1956]), though if α is not even aware in advance which random variable I will be revealed by information good $\mathbf{I}$, then we would further need to assume a prior over the process generating $\mathbf{I}$ and take the expetation over it $\mathbf{E}_{\mathbf{I} \sim \mathbf{P}[\mathbf{I}]}[\mathbf{E}_{i \sim I}[\mathbf{E}[U(I \mid I = i)]]]$.

A naive information market with no inspection would cause buyers to bid (and thefore incentivize sellers to optimize for) the ex-ante value of information. The proposed framework in Weiss et al. [2024] would cause buyers to bid the ex-post value of information.

However, ex-post VOI is still not enough: while knowing $I = i$ (inspecting $\mathbf{I}$) provides *some* information on $U(\mathbf{I})$, it does not provide all of it; there may still be information asymmetry, because $U(\mathbf{I})$ is itself a random variable not completely determined by $\mathbf{I}$ (much like in ordinary goods markets where you can inspect the good but still have information asymmetry). For example, I could represent some factual claim, but you do not actually know if it is correct until some evidence I' is presented for it. You might say: surely the information (I, I') would be more valuable than I alone, so as long as the market presents this option, we're good? But it is also possible to imagine a situation where I is *false* but believable, and I' falsifies it – then the ex-post VOI of (I, I') is less than that of I alone, but the true utility of (I, I') is more than that of I alone.

One way to think of this is: Equation (1) creates a *new* decision problem for α, that of deciding whether to buy $\mathbf{I}$ – or which information to buy out of a list of offers. The set of information goods offered $\mathcal{I} = \{\mathbf{I}_1, \ldots \mathbf{I}_k\}$ is a new decision problem, with utilities of each choice now given by the (unknown/random, much like $U(x)$) true value-of-information $U(\mathbf{I})$. Thus we may be offered another set of information goods $\{\mathbf{I}_1^1, \ldots \mathbf{I}_k^1\}$ to help us with this decision, ad recursum. In general we have a sequence of decision problems $\mathcal{X}^{n+1} = \mathcal{P}(\mathcal{I}^n)$ where $\mathcal{I}^n := \{\mathbf{I}_1^n, \ldots \mathbf{I}_k^n\}$ are the information goods offered to us to help us decide $\mathcal{X}^n$, and each choice corresponds to choosing a subset of that information to help decide $\mathcal{X}^n$, where $\mathcal{X}^0 = \mathcal{X}$ and $\mathcal{X}^1 = \mathcal{P}(\mathcal{I}^0) = \mathcal{P}(\mathcal{I})$.

There are two ways that we can set up these recursive problems. The naive approach, which arises from straightforwardly "applying Weiss et al. [2024] to itself", is to model each decision problem as having its own utility function $U^n : \mathcal{X}^n \to \mathbb{R}$ based on its instrumental utility for the previous decision problem $\mathcal{X}^{n-1}$. The problem with this approach is it fails to account for the possibility that a choice x^n can directly (i.e. not through its impact on x^{n-1}) impact a choice x^m where $m < n - 1$. This approach and an example illustrating why it is insufficient, are given in Appendix A.

[2] Including i in the tuple is to simplify notation when talking about taking conditional expectations on I. It is not necessary: instead of writing $\mathbf{E}[U(x) \mid I = i]$ we can think of $\mathbf{E}[U(x) \mid I]$ as itself a random variable correlated with I

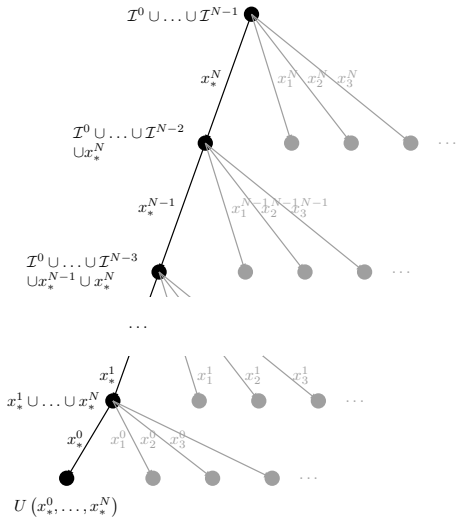


Figure 1: Recursive Inspection as an imperfect recall game; nodes are labelled by the information available for making the decision at that node. Note how the decision tree is in the reverse order of the inspection order: x^N is decided first, and x^0 last.

Instead our approach, the *recursive information protocol* implemented in Section 3, allows the agent (or rather the LLM subcontracted by the agent) to retain the full sequence of information bought in the recursive steps $x_*^{n+1}, \ldots x^N$ while making the decision $x^n \in \mathcal{X}^n$ (where N is some pre-defined finite depth we recurse till); furthermore x^n is decided keeping in mind the full traceback of decision problems $\mathcal{X}^0, \ldots \mathcal{X}^{n-1}$ that may be influenced by this decision. This is naturally modelled as an *imperfect recall game*[3] [Tewolde et al., 2024] where we first decide $x_*^N \in \mathcal{X}^N$ with full information $\mathcal{I}^0 \cup \ldots \mathcal{I}^{N-1}$, then $x_*^{N-1} \in \mathcal{X}^{N-1}$ with information $\mathcal{I}^0 \cup \ldots \mathcal{I}^{N-2} \cup x_*^N$ and so on until we finally decide $x_*^0 \in \mathcal{X}^0$ with information $x_*^1 \cup \cdots \cup x_*^N$. This is shown in Figure 1.

A node $(x^n, \ldots x^N)$ corresponds to the state where the agent has purchased $(x^n, \ldots x^N)$ and is choosing some $x^{n-1} \in \mathcal{X}^{n-1}$. For a Bayesian agent, we can thus recursively give the "value of being at a node":

$$U(x^n, \ldots x^N) = U\left(\arg\max_{x \in \mathcal{X}^{n-1}} \mathbf{E}\left[U(x, x^n, \ldots x^N) \mid \mathcal{I}^0 \cup \cdots \cup \mathcal{I}^{n-2} \cup x^n \cup \cdots \cup x^N\right], x^n, \ldots x^N\right) \tag{3}$$

$$U(x^0, \ldots x^N) = U(x^0) - \sum_{n=1}^{N} \sum_{\langle I,i,p \rangle \in x^n} p \tag{4}$$

This then completely specifies the behavior of a Bayesian agent performing a depth-N recursive inspection:

$$x_*^n = \arg\max_{x \in \mathcal{X}^n} \mathbf{E}\left[U(x, x^{n+1}, \ldots x^N) \mid \mathcal{I}^0 \cup \cdots \cup \mathcal{I}^{n-1} \cup x^{n+1} \cup \cdots \cup x^N\right] \tag{5}$$

In what sense is this algorithm "optimal"? It is instructive to first consider what notions of optimality *don't* hold for the recursive inspection protocol:

Non-theorem 2.1. *We might think that the resulting sequence $(x_*^0, \ldots x_*^N)$ is optimal given all the information present in the system $\mathcal{I}^0, \ldots \mathcal{I}^{N-1}$, i.e.*

$$\mathbf{x}_* = \arg\max_{\mathbf{x} \in \prod \mathcal{X}} \mathbf{E}\left[U(\mathbf{x}) \mid \mathcal{I}^0 \cup \cdots \cup \mathcal{I}^{N-1}\right]$$

[3]The decision-theoretic considerations in imperfect recall games do not matter to us, as the agent's actions themselves are not being forgotten — only the information offers

Counter-example. It is always best to simply buy x_*^0 and none of the subsequent x_*^ns that facilitated this decision. The same argument applies at any level, thus we cannot even make any claim like "$\mathbf{x}_*$ is optimal given the information purchased". □

Instead, our algorithm seems to be optimal in a "bounded rationality" sense: optimal in a way that also accounts for the *costs* of acquiring the information that would help improve our decision. One way to phrase this is: ex-ante (not knowing the information it is about to be offered), an agent would prefer to use this protocol compared to any other protocol.

Definition 2.2 (Admissible purchase protocol)**.** an "admissible purchase protocol" is a list of functions ξ^n mapping a decision problem $\mathcal{X}^0$ and a sequence of information offer sets $\mathcal{I}^0, \dots \mathcal{I}^{N-1}$ generated from it, to $\mathcal{X}, \mathcal{P}\left(\mathcal{I}^0\right), \dots \mathcal{P}\left(\mathcal{I}^{N-1}\right)$:

$$\begin{aligned} x^N &= \xi^N(\mathcal{I}^0, \dots \mathcal{I}^{N-1}) \\ &\dots \\ x^n &= \xi^n(\mathcal{I}^0, \dots \mathcal{I}^{n-1}, x^{n+1}, \dots x^N) \\ &\dots \\ x^0 &= \xi^0(x^1, \dots, x^N) \end{aligned}$$

i.e. a decision cannot "steal" information offers specifically made to help improve that decision, it can only depend on *purchased* information offers. We denote the tupled function as $(x^0, \dots x^N) = \xi(\mathcal{I}^0, \dots \mathcal{I}^{N-1}) = \xi(\mathcal{I})$.

In order to express "ex-ante expected utility", the agent must have a prior on the $\mathcal{I}^n$ that will be generated — this can be achieved either by having a measurable map $\Omega \to \text{finset}(\textsc{InfoOffers})^N$ (where $\textsc{InfoOffers} = \mathcal{F} \times \{0,1\} \times \mathbb{R}$ is the type of information goods) or more simply by assuming the random variables I_k^n revealed are fixed and only having a prior over their values. Both allow us to take an expectation over $\mathcal{I}^0, \dots \mathcal{I}^{N-1}$.

Theorem 2.3 (Recursive Inspections are ex-ante superior to any admissible protocol)**.** *Let x_*^n be the protocol described in Equations (3) to (5). Then for any admissible purchase protocol ξ:*

$$\mathbf{E}_{\mathcal{I}^0, \dots \mathcal{I}^{N-1}}\left[U(\xi(\mathcal{I}^0, \dots \mathcal{I}^{N-1}))\right] \leq \mathbf{E}_{\mathcal{I}^0, \dots \mathcal{I}^{N-1}}\left[U(x_*(\mathcal{I}^0, \dots \mathcal{I}^{N-1}))\right]$$

Proof. Let ξ_{*n} denote the following admissible protocol:

$$\begin{aligned} x^N &= \xi^N(\mathcal{I}^0, \dots \mathcal{I}^{N-1}) \\ &\dots \\ x^{n+1} &= \xi^{N+1}(\mathcal{I}^0, \dots \mathcal{I}^n, x^{n+2}, \dots x^N) \\ x^n &= x_*^n(\mathcal{I}^0, \dots \mathcal{I}^{n-1}, x^{n+1}, \dots x^N) \\ &\dots \\ x^0 &= x_*^0(x^1, \dots, x^N) \end{aligned}$$

Then $\xi_{*N} = x_*$ and $\xi_{*-1} = \xi$. It suffices to show that ξ_{*n+1} dominates ξ_{*n}, i.e.

$$\mathbf{E}_{\mathcal{I}^0, \dots \mathcal{I}^{N-1}}\left[U(\xi_{*n}(\mathcal{I}^0, \dots \mathcal{I}^{N-1}))\right] \leq \mathbf{E}_{\mathcal{I}^0, \dots \mathcal{I}^{N-1}}\left[U(\xi_{*n+1}(\mathcal{I}^0, \dots \mathcal{I}^{N-1}))\right]$$

Observe that (where we have omitted the parameters to x_*^n, ξ^n etc. as a shorthand):

$$\begin{aligned} U(\xi_{*n}(\mathcal{I}^0, \dots \mathcal{I}^{N-1})) &= U\left(x_*^0, \dots x_*^n, \xi^{n+1}, \dots \xi^N\right) \\ &= U\left(\xi^{n+1}, \dots \xi^N\right) \end{aligned}$$

Where we have recursively applied Equation (3) to transform the expression into one about the value of a node at level $n+1$. Thus the goal is reduced to:

$$\mathbf{E}_{\mathcal{I}^0, \dots \mathcal{I}^{N-1}}\left[U(\xi^{n+1}, \dots \xi^N)\right] \leq \mathbf{E}_{\mathcal{I}^0, \dots \mathcal{I}^{N-1}}\left[U(x_*^{n+1}, \xi^{n+1}, \dots \xi^N)\right]$$

From Equation (5) we have that this is true for the expectation over $\mathcal{I}^0, \dots \mathcal{I}^n$; we can then simply take the expectation over $\mathcal{I}^{n+1}, \dots \mathcal{I}^{N-1}$ and have our result. □

Algorithm 1 One-level Inspection Protocol from Weiss et al. [2024]

class BUYERCONTEXT
  $\mathcal{X}$: decision problem it wants information for
  $D : \text{list}[\text{str}] \to \mathcal{X}$, decision procedure based on available information
end class
class SELLER
  A : BUYERCONTEXT $\to \text{str} \times \mathbb{R}$, generate INFOOFFER and price
end class
procedure IP(Q : BUYERCONTEXT)
  ▷ Post contextual information to sellers to receive INFOOFFERS
  $\mathcal{I} \leftarrow \{\beta(Q) \text{ for } \beta \in \text{SELLERS}\}$
  ▷ Use an LLM to decide which **I** to buy
  $I^* \leftarrow \texttt{LLM}(\texttt{prompt} = \text{“You need to buy an } \mathbf{I} \text{ from } \mathcal{I} \text{ to help decide } Q\text{”})()$
  ▷ Decide based on purchased information
  $x^* \leftarrow Q.D(I^*)$
  return x^*
end procedure

Algorithm 2 Recursive Inspection Protocol

procedure RIP(Q : BUYERCONTEXT)
  ▷ Post to sellers and get INFOOFFERS from them
  $\mathcal{I} \leftarrow \{\beta(Q) \text{ for } \beta \in \text{SELLERS}\}$
  ▷ Create Recursive BuyerContext to help decide Q
  $Q' \leftarrow \text{BUYERCONTEXT}(\mathcal{I}, \texttt{LLM}(\texttt{prompt} = \text{“You need to buy an } \mathbf{I} \text{ from } \mathcal{I} \text{ to help decide } Q\text{”}))$
  ▷ Get INFOOFFERS chosen in Q'
  $I^* \leftarrow \text{RIP}(Q')$
  ▷ Decide based on all collected information from recursive steps
  $x^* \leftarrow Q.D(I^*)$
  return x^*, I^*
end procedure

3 Recursive Inspection Protocol

Algorithm 1 shows the inspection protocol for information markets introduced in Weiss et al. [2024]: information sellers offer information goods to a buyer, who spins off an LLM to inspect and purchase the information goods. Our *Recursive Inspection Protocol* extends this by letting the subcontracted LLM buyer further consult the information market (spin off another sub-LLM) to help it make its decision, ad recursum. A simplified version of the logic, ignoring server implementation details, is presented in in Algorithm 2 and Figure 2.

A working implementation of an information market server implementing the Recursive Inspection Protocol is available at the `infonomy-server` repository[4] – some screenshots of the platform's GUI are shown in Figure 3. Many applications of such a server are immediate:

- *Question & Answer site* – As presented, `infonomy-server` can be seen as a Q & A site with market incentives for answering questions.
- *Privatized product regulation* – The BUYERCONTEXTS might be names or links to products (the decision being “should I buy?”), and the INFOOFFERS could be inspection checks done by private labs, or customer reviews, which are now incentivized to be answerable to the customer.
- *Community Notes* – In the spirit of well-known crowdsourced fact-checking systems such as Community Notes/Birdwatch [Wojcik et al., 2022], we might use information markets as a “comments section for the internet”, where BUYERCONTEXTS would be links to webpages or social media posts (the decision being “should I believe?”) and the INFOOFFERS would be fact-checks or important context.

[4] https://anonymous.4open.science/r/infonomy-server-5668/

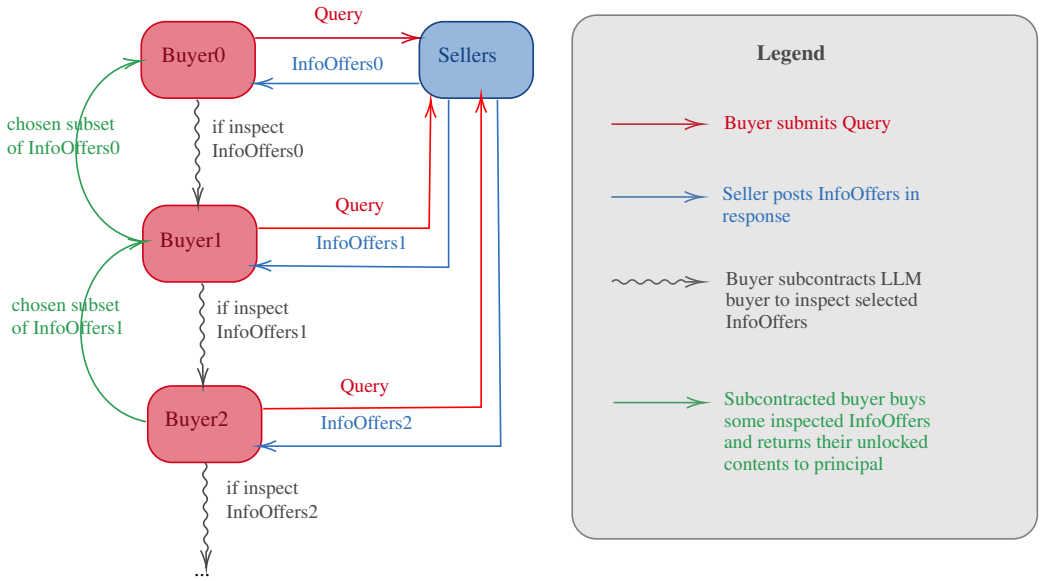
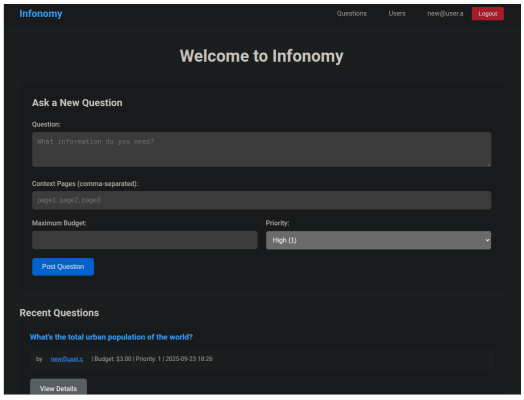
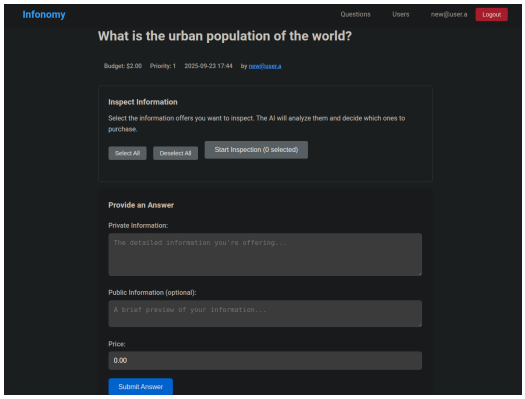


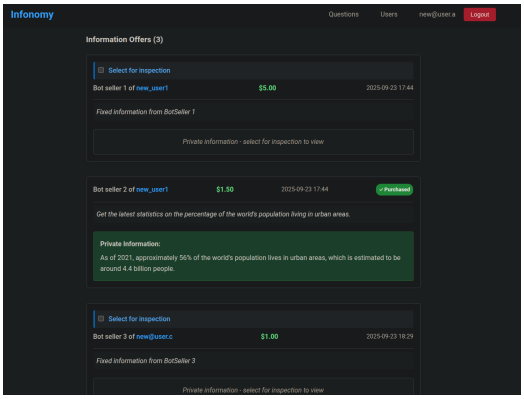
Figure 2: The Recursive Inspection Protocol



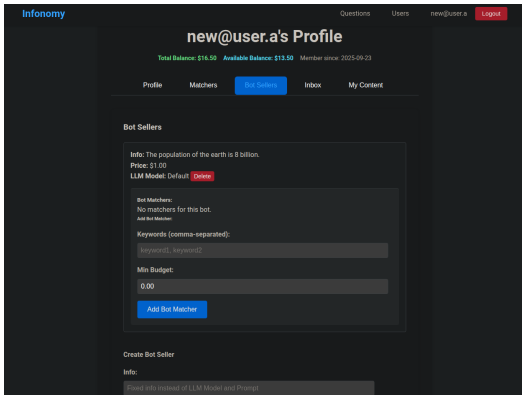
(a) Posting a new question (BUYERCONTEXT); viewing BUYERCONTEXTSon the server



(b) Posting an answer (INFOOFFER); initiating a recursive inspection



(c) Inspecting and purchasing INFOOFFERS; viewing purchased INFOOFFERS



(d) Bot sellers automatically answer recursive BUYERCONTEXTS

Figure 3: Screenshots from the `infonomy-server` platform

- *Reasoning in prediction markets* – Forecasters on prediction markets may benefit from incentivizing the provision of information relevant to a forecast. One solution to this was given by Conitzer [2009]; another is to use `infonomy-server`: where the BUYERCONTEXTS are the questions being forecasted on ("What is the correct probability of this question?") and the INFOOFFERS are any relevant pieces of information.

4 Conclusion

We have introduced the *recursive inspection protocol* as an information market mechanism for agents that are capable of "inspecting and forgetting" information, and provided a working implementation of the algorithm employing LLM agents. Our mechanism is optimal in a certain fundamental sense that accounts for the cost of obtaining information.

The most promising direction for future work is to explore applications of this mechanism to *scalable oversight* in AI alignment. Our mechanism may be seen as a formalization of Extrapolated Volition [Yudkowsky, 2004], which is informally described as the "action an agent would prefer if it had all the relevant information". Algorithm 2 itself may be remniscient of scalable oversight proposals such as "Humans Consulting HCH" [Christiano, 2018].

More generally, one of the key limitations of currently widely-used alignment techniques such as reinforcement learning from human feedback (RLHF) is that they fundamentally rely on a human's ability to judge the correctness or value of a (potentially superhuman) AI's outputs [Casper et al., 2023, Burns et al., 2024]. In other words, the AI is trained on the human supervisor's *immediate, superficial* volition, rather than on her *extrapolated volition* [Yudkowsky, 2004]. In this context, our proposal may be used to augment a human rater's judgement of the AI's outputs with an information market, better reflecting her "extrapolated volition".

References

George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier, 1978.

K. J. Arrow. *Economic Welfare and the Allocation of Resources for Invention*, pages 219–236. Macmillan Education UK, London, 1972. ISBN 978-1-349-15486-9. doi: 10.1007/978-1-349-15486-9_13. URL https://doi.org/10.1007/978-1-349-15486-9_13.

Moshe Babaioff, Robert Kleinberg, and Renato Paes Leme. Optimal mechanisms for selling information. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, page 92–109, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314152. doi: 10.1145/2229012.2229024. URL https://doi.org/10.1145/2229012.2229024.

Yoram Barzel. Transaction Costs: Are They Just Costs? *Zeitschrift für die gesamte Staatswissenschaft / Journal of Institutional and Theoretical Economics*, 141(1):4–16, 1985. ISSN 00442550.

Michael Ben-Or, Oded Goldreich, Shafi Goldwasser, Johan Håstad, Joe Kilian, Silvio Micali, and Phillip Rogaway. Everything provable is provable in zero-knowledge. In Shafi Goldwasser, editor, *Advances in Cryptology – CRYPTO 1988 - Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 37–56, Germany, 1990. Springer Verlag. ISBN 9780387971964. doi: 10.1007/0-387-34799-2_4. Publisher Copyright: © Springer-Verlag Berlin Heidelberg 1990.; Conference on Theory and Applications of Cryptography, CRYPTO 1988 ; Conference date: 21-08-1988 Through 25-08-1988.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The Design and Price of Information. *The American Economic Review*, 108(1):1–48, 2018. ISSN 0002-8282.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova

DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring Progress on Scalable Oversight for Large Language Models, November 2022.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4971–5012. PMLR, July 2024.

Vitalik Buterin. From prediction markets to info finance. https://vitalik.eth.limo/general/2024/11/09/infofinance.html, 2024. URL https://vitalik.eth.limo/general/2024/11/09/infofinance.html. Accessed: October 9, 2025.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Ségerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca D. Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Junjie Chen, Minming Li, and Haifeng Xu. Selling Data To a Machine Learner: Pricing via Costly Signaling. In *Proceedings of the 39th International Conference on Machine Learning*, pages 3336–3359. PMLR, June 2022.

Paul Christiano. Humans Consulting HCH. https://ai-alignment.com/humans-consulting-hch-f893f6051455, November 2018. URL https://ai-alignment.com/humans-consulting-hch-f893f6051455.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts, 2018. URL https://arxiv.org/abs/1810.08575.

Vincent Conitzer. Prediction markets, mechanism design, and cooperative game theory. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 101–108, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism Design for Large Language Models. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, pages 144–155, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645511.

Alireza Fallah, Michael Jordan, Ali Makhdoumi, and Azarakhsh Malekian. On three-layer data markets. *ArXiv*, abs/2402.09697, 2024. URL https://api.semanticscholar.org/CorpusID:267682401.

Benja Fallenstein and Nate Soares. Vingean reflection: Reliable reasoning for self-improving agents. Technical Report 2015-2, MIRI, 2015. URL https://intelligence.org/files/VingeanReflection.pdf.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

O. Goldreich. *Foundations of Cryptography: Volume 1, Basic Tools*. Cambridge University Press, 2001. ISBN 9780521791724. URL https://books.google.co.uk/books?id=oo3RzgEACAAJ.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching Human-Level Forecasting with Language Models, February 2024.

Lewis Hammond and Sam Adam-Day. Neural Interactive Proofs. In *ICML 2024 Next Generation of AI Safety Workshop*, July 2024.

Robin Hanson. IP+ Like Barbed Wire?, July 2011a.

Robin Hanson. Rah Efficient IP, July 2011b.

R. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1): 22–26, 1966. doi: 10.1109/tssc.1966.30007.

Evan Hubinger. Alignment proposals and complexity classes. https://www.lesswrong.com/posts/N64THGX7XNCqRtvPG/alignment-proposals-and-complexity-classes, 2020. URL https://www.lesswrong.com/posts/N64THGX7XNCqRtvPG/alignment-proposals-and-complexity-classes. Accessed: October 9, 2025.

Russell Impagliazzo and Moti Yung. Direct minimum-knowledge computations. In *A Conference on the Theory and Applications of Cryptographic Techniques on Advances in Cryptology*, CRYPTO '87, page 40–51, Berlin, Heidelberg, 1987. Springer-Verlag. ISBN 3540187960.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, October 2018.

H. W. Kuhn, K. J. Arrow, E. W. Barankin, D. Blackwell, R. Bott, N. Dalkey, M. Dresher, D. Gale, D. B. Gillies, I. Glicksberg, O. Gross, S. Karlin, H. W. Kuhn, J. P. Mayberry, J. W. Milnor, T. S. Motzkin, J. von Neumann, H. Raiffa, L. S. Shapley, M. Shiffman, F. M. Stewart, G. L. Thompson, and R. M. Thrall. *Extensive games and the problem of information*, pages 193–216. Princeton University Press, 1953. ISBN 9780691079356. URL http://www.jstor.org/stable/j.ctt1b9x1zv.17.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.829.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. ISSN 00034851, 21688990. URL http://www.jstor.org/stable/2237191.

Alejandro M. Manelli and Daniel R. Vincent. Bundling as an optimal selling mechanism for a multiple-good monopolist. *Journal of Economic Theory*, 127(1):1–35, 2006. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2005.08.007. URL https://www.sciencedirect.com/science/article/pii/S0022053105002395.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency Checks for Language Model Forecasters. In *The Thirteenth International Conference on Learning Representations*, October 2024.

P.A. Samuelson and W.D. Nordhaus. *Economics*. McGraw-Hill Education, 2009. ISBN 9780073511290. URL https://books.google.co.uk/books?id=eS5ZAAAAYAAJ.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy, May 2024.

George J Stigler. The economics of information. *Journal of political economy*, 69(3):213–225, 1961.

Emanuel Tewolde, Brian Hu Zhang, Caspar Oesterheld, Manolis Zampetakis, Tuomas Sandholm, Paul Goldberg, and Vincent Conitzer. Imperfect-recall games: equilibrium concepts and their complexity. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/332. URL https://doi.org/10.24963/ijcai.2024/332.

Kristine Thomassen, Siril Vassbø, Espen Solheim-Kile, and Jardar Lohne. Public-private partnership: Transaction costs of tendering. *Procedia Computer Science*, 100:818–825, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.09.230. URL https://www.sciencedirect.com/science/article/pii/S1877050916323997. International Conference on ENTERprise

Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016.

Marshall V. Van Alstyne. A proposal for valuing information and instrumental goods. In *Proceedings of the 20th International Conference on Information Systems*, ICIS '99, pages 328–345, USA, January 1999. Association for Information Systems.

Martin Weiss, Nasim Rahaman, Manuel Wuthrich, Yoshua Bengio, Li Erran Li, Bernhard Schölkopf, and Christopher Pal. Redesigning Information Markets in the Era of Language Models. In *First Conference on Language Modeling*, August 2024.

Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation, 2022. URL https://arxiv.org/abs/2210.15723.

Eliezer Yudkowsky. Coherent Extrapolated Volition, 2004.

A Successive Inspection Protocol

We describe here the naive version of the recursive inspection protocol, or "simply applying Weiss et al. [2024] to itself".

Table 1: *Successive decision problems, naive approach*

Choice set	**True utilities**	**Information Offers**
$\mathcal{X}$	$U : \mathcal{X} \to \mathbb{R}$	$\mathcal{I} = \{\mathbf{0}, \mathbf{I}_1, \mathbf{I}_2, \dots\}$
$\mathcal{X}^1 = \mathcal{I}$	$U^1 : \mathcal{X}^1 \to \mathbb{R}$	$\mathcal{I}^1 = \{\mathbf{0}, \mathbf{I}_1^1, \mathbf{I}_2^1, \dots\}$
$\mathcal{X}^2 = \mathcal{I}^1$	$U^2 : \mathcal{X}^2 \to \mathbb{R}$	$\mathcal{I}^2 = \{\mathbf{0}, \mathbf{I}_1^2, \mathbf{I}_2^2 \dots\}$
$\dots$	$\dots$	$\dots$

$$U^{n+1}\left(\langle I, i, p\rangle\right) := U^n\left(\arg\max_{x \in \mathcal{X}^n} \mathbf{E}\left[U^n(x) \mid I = i\right]\right) - U^n\left(\arg\max_{x \in \mathcal{X}^n} \mathbf{E}\left[U^n(x)\right]\right) - p \quad (6)$$

These choice sets and utilities are very similar to the decision problems created by the *recursive inspection protocol* described in Algorithm 2 and in the main body; except that each action $x^n \in \mathcal{X}^n$ is made only consulting the information chosen in $x^{n+1} \in \mathcal{X}^{n+1}$. We can then say that the decisions made under this protocol are:

$$x_*^n = \arg\max_{x \in \mathcal{X}^n} \mathbf{E}\left[U^n(x) \mid x_*^{n+1}\right] \quad (7)$$

In general this is an ill-defined infinite recursion. But finite restrictions of this are natural, stopping at some fixed $x_{*:N}^N = \arg\max_{x \in \mathcal{X}_N} \mathbf{E}\left[U^N(x)\right]$ (you can think of this stopping as caused by the transaction costs of inspection), so that for $n < N$:

$$x_{*:N}^n = \arg\max_{x \in \mathcal{X}^n} \mathbf{E}\left[U^n(x) \mid x_{*:N}^{n+1}\right] \quad (8)$$

Counter-example. Take the decision problem $\mathcal{X}^0 = \{x_0, x_1, x_2\}$ where $x_0 =$ eat raw legume, $x_1 =$ eat rice and $x_2 =$ eat boiled legume. Then in reality, $U(x_2) > U(x_1) >> U(x_0)$ (rice is unhealthy, but raw legumes are toxic).

Suppose $\mathcal{I}^0 = \{I_0^1, I_1^1\}$ where $I_0^1 =$ legumes are toxic, $I_1^1 =$ rice is unhealthy and $I_0^1 =$ legumes are toxic; and $\mathcal{I}^1 = \{I_0^2\}$ where $I_0^2 =$ the toxins in legumes can be removed by boiling. All of these information offers could even be free.

Then the optimal action is $(x_2, \{I_0^1, I_1^1\}, \{I_0^2\})$; however, if the information bought in level-2 $\{I_0^2\}$ is not available while deciding x^0, the best the agent can do is $(x_2, \{I_0^1\}, \{I_0^2\})$ to prevent itself from eating raw legumes (since it will not know that the toxins can be removed by boiling). □

B Basic results about value-of-information

We include two basic lemmas which are perhaps obvious, but upon which the whole theory of value-of-information rests. Lemma B.1 asserts "A Bayesian agent expects to gain from information"[5]. Lemma B.2 asserts: "A Bayesian agent expects to gain from inspection"[6] (simply because it can always choose not to buy anything after inspecting), thus in the absence of transaction costs of inspection buyers *will* want to inspect more and more at each step.

Lemma B.1 (Bayesian agent expects to gain from information)**.** *Assume a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, a set of choices $\mathcal{X}$ and a measurable utility function $U : \Omega \times \mathcal{X} \to \mathbb{R}$. Then for any random variable $I : \Omega \to \mathbb{R}$, we have:*

$$\mathbf{E}_I \left[U(\arg\max_x \mathbf{E}\left[U(x) \mid I\right]) \right] \geq \max \mathbf{E}\left[U(x)\right]$$

Proof. For all values of $I = i$ and for all $x_0 \in \mathcal{X}$, we have (from the definition of $\arg\max$):

$$\mathbf{E}\left[U(\arg\max_x \mathbf{E}\left[U(x) \mid I = i\right]) \mid I = i \right] \geq \mathbf{E}\left[U(x_0) \mid I = i\right]$$

We take the expectation over $i \sim I$ on both sides, apply the law of total expectation and set $x_0 = \arg\max \mathbf{E}\left[U(x)\right]$ to obtain our result. □

Lemma B.2 (Bayesian agent expects to gain from inspection)**.** *Let the recursive construction of $\mathcal{X}^n$, U^n, $\textsc{InfoOffers}^n$,x_*^n be as in Section 2. Then $\forall n$, $\mathbf{E}\left[U^{n+1}(x_*^{n+1})\right] \geq 0$. The same also applies to finite-depth inspections i.e. $\mathbf{E}\left[U^{n+1}(x_{*:N}^{n+1})\right] \geq 0$.*

Proof. Equation (7) implies that $\mathbf{E}\left[U^n\left(x_* \mid x_*^{n+1}\right)\right]$ is $\geq$ any version of the same expression with x_*^n replaced by any $x \in \mathcal{X}^n$. We choose $x = \mathbf{0}$, for which the expression evaluates to 0, and take the expectation over x_*^{n+1}. □

[5] From an AI alignment perspective, this may instead be phrased "A Bayesian agent trusts a more informed version of itself", and is a statistical-information version of *Vingean reflection* or *tiling agents* [Fallenstein and Soares, 2015], the desire that (even logically non-omniscient) agents can trust smarter versions of themselves.

[6] Lemma B.2 is also the corrected and generalized version of an incorrectly formulated result in the arXiv version of Weiss et al. [2024]