Reinforcement Learning from Recursive Information Market Feedback

# Summary

RLHF is no good on tasks which humans are unable to easily “rate” output. I propose the *Recursive Information Market*, which can be understood as an approach to rate based on a human rater’s Extrapolated Volition, or a generalized form of AI safety via debate.

# The non-summary

## Introduction

One way to frame the 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 (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).

One natural framework that comes to mind for determining the “true” (or *best estimated*) value of some good (in this case information provided by an AI) is *markets[[1]](#footnote-0)*. One would naively expect that simply implementing a marketplace where AIs sell information to human buyers would result in the highest-value (in particular, *true*) information being supplied. There are two reasons this does not occur:

• **Information markets are inefficient.** Information markets suffer from the *buyer’s inspection paradox* – 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.

• **Markets are only efficient in the absence of information asymmetry (Barzel, 1985).** Information asymmetry here includes computational asymmetries, which are most relevant for our purposes[[2]](#footnote-1) – if the human buyer does not know if the information supplied by the AI is correct, then his buying decision is indifferent to its correctness, and the AI is not incentivized to provide truthful or valuable informaton.

If not for the first problem, we could have hoped to address the second problem could by simply buying purchase-relevant information on a second market – alas, the first point tells us that information markets do not work in the first place.

Rahaman et al. (2024) suggests a novel solution to the buyer’s inspection problem: *just have an LLM make the purchase decision on your behalf!* LLMs, unlike humans, are capable of simply “forgetting” (deleting from the context window) some information they see in inference-time; thus they can inspect information and reliably commit to not stealing it.

I propose: the ***Recursive Information Market***, an extension of the mechanism in Rahaman et al. (2024) to address the problem of making purchases in the presence of information asymmetry, such as when buying information from a superhuman AI.

## Recursive Information Market

The overall picture is as follows: suppose an AI *A*0 gives us some answer *x*0 to a query; there may be many pieces of information which may influence how much you (or really your LLM representative *B*0) are willing to pay for this information – so your LLM representative spins off *another* LLM representative[[3]](#footnote-2) *B*1 to buy information *x*1 from a secondary information market of AI informants *A*1. But, crucially, *B*1 might also not be able to confidently verify the quality of *x*1, so it can spin-off another LLM representative, and so on. Each spin-off incurs some transaction cost paid by the respective AI informant *An*, and the process terminates when the value of the information no longer justifies the transaction cost[[4]](#footnote-3).

Here’s a pseudo-pythonic sketch of the algorithm:

**class** Buyer :

goal : str

wealth : **float**

info\_processing\_cost : **float** *# could be a function instead*

**def** \_\_call\_\_ ( self ) -> tuple [ list [ str], **bool** ]:

*# initialize info\_collected*

info\_collected = []

*# tell information agents your goals and get info offers from them* info\_offers = Arena . offer\_info [ self ]

top\_offer = max( info\_offer , key= **lambda** offer , offer . bid )

**if** top\_offer . bid > info\_processing\_cost :

*# charge winning advertiser for cost of considering info*

self . wealth += top\_offer . bid

top\_offer . parent . wealth -= top\_offer . bid

*# spin off contractor agent to decide if to buy top offer*

contractor = Buyer (

goal = DecideToBuy ( top\_offer ) ,

wealth = self . wealth ,

info\_processing\_cost = self . info\_processing\_cost

)

info\_collected , decision = contractor ()

**if** decision :

*# buy the info*

info\_collected . append ( top\_offer . info )

self . wealth -= top\_offer . price

top\_offer . parent . wealth += top\_offer . price

**return** info\_collected , self . decide ( info = info\_collected )

**def** decide ( self , info : list [str ]):

... *# some intelligent behaviour*

**class** Informer :

wealth : **float**

@dataclass

**class** InfoOffer :

bid: **float**

price : **float**

info : str

parent : Informer

@property

**def** offer\_info ( self ) -> dict [Buyer , InfoOffer ]:

*# some daemon that monitors the arena for places it could*

*# be useful , and advertises its info there*

...

**class** Arena :

buyers : list [ Buyer ]

informers : list [ Informer ]

@property

**def** offer\_info ( self ):

info\_offers = **{** buyer : [] **for** buyer **in** self . buyers **}**

**for** informer **in** self . informers :

**for** buyer **in** info\_offers :

info\_offers [ buyer ]. append ( informer . offer\_info )

**return** info\_offers

In this mechanism as described, the AI we’re trying to align *A*0, as well as all the subsequent “informants” *A*1, *A*2, *. . .* are expected to be superhuman, whereas the “LLM representatives” employed are intended to be current, approximately human-level LLMs. In essence the core of our proposal tries to solve the problem of **aligning superhuman AIs *to* current human-level LLMs**, rather than to humans directly. See #Limitations for some discussion on this.

## Project plan

Three specific/“actionable” lines of work I expect will be central in this project:

### 1) Theoretical and foundational work / expected results

We want:

1. A **precise economic model** with information asymmetry, demonstrating how and *when* the Recursive Information Market results in greater economic efficiency/surplus.
2. The “when” in the above would be formulated in a rather abstract way — in a specific task setting like “[AI systems performing computations, where each step in the computation can be judged by humans](https://www.lesswrong.com/posts/DGt9mJNKcfqiesYFZ/debate-oracles-and-obfuscated-arguments-3)”, we could better **characterize this as a complexity class**, similar to in the debate paper. This way we could demonstrate that our proposal is “at least as good as Debate”.
3. I could also imagine a world in which humans were literally incapable of judging an AI’s outputs no matter how much info you gave them. In such a world — or rather for *such tasks/problems*, which are outside the above “complexity class” — the value of the AI’s outputs should approach 0, or rather the maximum level that a human is capable of judging, and the human shouldn’t buy anything more than that, resembling the decay of trust in an economy. In other words, we would like to demonstrate that **for problems outside the complexity class, our mechanism “plays safe”** — and also compare against how Debate behaves in such cases.
4. If we model the LLM buyer as a Bayesian agent (with utility function dictated by prompt) lacking only in real “statistical” information compared to the superhuman AI, then the desired result in I would be quite straightforward. Otherwise, the buyer would have predictable biases that the AI could exploit. However, we might wonder: ***what kinds of belief-update behaviors can be “rationalized” as Bayesian so that the results still apply?*** We would consider this a definition of “bounded rationality”. We might not be able to produce a full characterization, but some examples would be nice — in particular, I believe that a converse market of buyers, each endowed with some wealth and competing on some ground truth task (e.g. betting on prediction markets, or producing economically valuable work), would be a boundedly rational agent.
5. We would like to compare our mechanism to other Scalable Oversight protocols. There appears to be a very strong analogy to **AI Safety via market-making** (Evan Hubinger, 2020) as well as the “Information agents and Probability agents” mechanism introduced at the end of Conitzer (2009); crucially I think our mechanism produces better incentives for the information agents, but this should be proven.
6. Another promising direction might be to show an **equivalence between our framework and a co-operative game between the AI informants**– this is motivated by Conitzer (2009), which demonstrates that this is true for certain information market mechanisms.

### 2) Implementations in other contexts

Nothing in the mechanism we’ve outlined is restricted to *buying information from superhuman AIs*. The implications of our idea are quite far-reaching, and can be applied to any situation where you buy information, or you buy goods under information asymmetry:

* Intellectual property
* Positive externalities of prediction markets
* Information asymmetry in real markets
* Fact-checking/community notes
* Bounties for tasks

(see my [LessWrong post](https://www.lesswrong.com/posts/Y79tkWhvHi8GgLN2q/reinforcement-learning-from-information-bazaar-feedback-and) for some explanation)

Even though our main focus will be to improve RLHF, I have no idea if we’d be able to actually use it in training – and honestly I don’t think we should until we’ve tested it in these other, less risky contexts: these are all valuable subprojects to implement.

### 3) Evaluation as a Scalable Oversight Protocol

Our algorithm can generally be considered a scalable oversight protocol. Even if we can’t actually train with RLRIMF, we can do [“inference-only” experiments](https://www.lesswrong.com/posts/ot25KcanipSaKzAH9/inference-only-debate-experiments-using-math-problems-1) to empirically evaluate its performance as a scalable oversight protocol.

## Limitations

**1)** As it stands, the core of our proposal tries to solve the problem of **aligning superhuman AIs *to* current human-level LLMs**, rather than to humans directly. This is just a matter of “happenstance”, resulting from the fact that LLMs are able to forget while human buyers aren’t, so I would expect that this to be in principle solvable, e.g. by training the buyer LLMs in parallel, perhaps based on their performance (or the performance of the humans rating them) on some ground truth task, e.g. forecasting.

A risk worth mentioning is that if LLM buyers are trained in tandem, they may learn to seek out deceptive (or “persuasive to humans, but not true”) information. Training based on some ground truth performance would solve this, but it’s not clear how this would work for tasks where ground truth is hard or will take a long time to get. Regardless, a protocol for aligning superhuman AIs with current LLMs is in itself a valuable contribution, because (1) it reduces the problem to aligning approximately current-level AIs with human principals, and (2) this remaining “alignment” problem is really a capabilities problem, of training LLMs that are good at pricing information – so this is something you would naturally expect to see advancement on.

**2)** Collusion between the AI informants is in principle possible, but is always made less likely by having a competitive market as opposed to something like Debate where the debaters are fixed a priori. **Ultimately the bare minimum I would like from this project is something that (provably) outperforms debate.**

**3)** My generalized picture of the AI alignment problem currently looks like this:

1. **Outer alignment** is the problem of making purchases under *information asymmetry*.
2. **Inner alignment** is the problem of “the reward function is not the utility function”.

Our proposal mainly addresses outer alignment.

# Output

The goal will be a paper (or multiple) + a github repo consisting of the algorithm and easy endpoints to integrate into RLHF as well as other applicable contexts.

# Risks and downsides (externalities)

What we’re proposing is a form of “[actual RL](https://x.com/karpathy/status/1821277264996352246)” (i.e. potentially not limited by human rater capacity) for LLMs. This is why I think it is critical to not actually run the algorithm until we have solid theoretical results about its safety.

With that said, the upside of our project vastly outweighs the risk, because we introduce no new risks:

* Debate is also “actual RL” and is being widely-researched. Any risks of our method also apply to Debate.
* AI companies are already training on things like agentic workflows with materially measurable reward. This is much riskier than our method because it optimizes for some very parochial rewards and can lead to paperclip maximizers.
* The idea is already out there now, and someone was bound to have come up with the idea sooner or later; it’s best we actually do theoretical work on it to make sure it goes well.

# Team

**Team size**

3-6 people including me.

**Research Lead**

Abhimanyu Pallavi Sudhir.

Email: [abhimanyupallavisudhir@gmail.com](mailto:abhimanyupallavisudhir@gmail.com)

My CV: <https://abhimanyu.io/cv/relevant.pdf>

I’ll be dedicating at least ~20h/wk to this project, and will likely be my main focus for the period of this project.

**Skill requirements**

Minimum skills:

* **Good economic intuitions:** I might ask some application questions of the form “Argue against this gut instinct: …”
* **General fluency in Python:** Writing code is one of the two main things we do to make sure the things we’re saying aren’t fake bullshit (the other way is betting on prediction markets). I would like to make sure that everyone is capable of writing clean code, and is generally “fluent” in stuff.

I would be especially happy to have someone who can make significant contributions on theoretical work, i.e. on coming up with and proving solid, useful theorems. Someone who can do the bulk of the effort on this I would happily grant joint-first-author position to.

The majority of team members, I assume, would be working on implementations in the various contexts stated earlier.

# References

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.

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.

Paul Christiano. Extracting information. https://ai-alignment.com/extracting-information-97cd956f2c17, October 2016.

Vincent Conitzer. Prediction Markets, Mechanism Design, and Cooperative Game Theory. In *Conference on Uncertainty in Artificial Intelligence*, June 2009.

Evan Hubinger. AI safety via market making — LessWrong, June 2020.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI Alignment from Human Feedback, May 2024.

Katja Grace. How to buy a truth from a liar, July 2014.

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

Nasim Rahaman, Martin Weiss, Manuel Wüthrich, Yoshua Bengio, Li Erran Li, Chris Pal, and Bernhard Schölkopf. Language Models Can Reduce Asymmetry in Information Markets, March 2024.

Eliezer Yudkowsky. Coherent Extrapolated Volition, 2004.

1. This is not a framing foreign to alignment, see e.g. Grace (2014) (“Buying truth from a liar”) and subsequent Christiano (2016). There has been a little discussion about applying mechanism design to RLHF, such as in the [Models of Human Feedback](https://sites.google.com/view/mhf-icml2024/accepted-papers?authuser=0) workshop at ICML, and [Axioms for AI alignment from Human Feedback by Ge et al](https://arxiv.org/abs/2405.14758), but it’s mostly focused on stuff like preference aggregation. [↑](#footnote-ref-0)
2. You might be willing to pay more for a more durable laptop, or for higher-polyphenolic olive oil, but you have no way of obtaining and verifying this information, so it does not affect your buying decision, and thus the supplier is incentivized to provide the cheapest olive oil he can find without regard for phenolic content. [↑](#footnote-ref-1)
3. by “another LLM representative”, we mean one with an independent context [↑](#footnote-ref-2)
4. In this description, an entire market of AI informants is trained simultaneously, taking their profits as training signal. This is not strictly necessary; however, I expect that this set-up is useful for preventing *collusion*. [↑](#footnote-ref-3)