Economic models of alignment and extrapolated volition

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

We propose a simple formal definition of AI alignment motivated by an economic analogy. Market failures (such as imperfect information and imperfect assurance) naturally correspond to misalignment modes: in particular, though we leave this for future work, we believe that a "recursive information markets" mechanism naturally leads us to a formalization of extrapolated volition in terms of a value-of-information expression.

1 Alignment as Pareto efficiency

Let $\alpha_1, \dots \alpha_n$ be some agents with action sets $X_1, \dots X_n$ and utility functions $U_1(\mathbf{x}), \dots U_n(\mathbf{x})$. In general if they are individually maximizing their utility functions, then maybe their chosen actions $\mathbf{x}^*$ will be some Nash-equilibrium of the game.

One question we can ask is: when can we model this system of two agents as a single "super-agent"? There are two natural ways to answer this question:

- When this Nash equilibrium $\mathbf{x}^*$ is the maximum of some "total utility function", where this total utility function has some sensible properties (like being increasing in each agent's utility)
- If the agents could co-ordinate and choose their action, there is no other joint action that would be better for all of them – i.e. is strongly Pareto-optimal.

In fact these notions are equivalent under some reasonable assumptions. Usually this is demonstrated in the case where the total utility function is a linear weighting of the agents' utilities [Negishi, 1960, Varian, 1976] although e.g. a previous Iliad paper [Little, 2025] also apparently showed that maximizing a "geometric" social welfare function is also Pareto-efficient.

The claim is then that **Pareto-optimality is a formal notion of "alignment" between the agents**.

Usually, one imagines alignment as a set-up where an agent β is constructed to have the *same* utility function as a human α (same in the sense of wanting the same allocation to α, not to itself). There are several intuitive reasons to prefer our "extrinsic" definition over this usual "intrinsic" one.

- **Incentive-compatible mechanisms.** In practical *existing* "alignment problems" in the world (e.g. aligning firms or markets), we do not solve them by totally reprogramming their values, but by making it profitable for them to do what we want. Therefore this frame is more useful for intuition-pumping from analogies.
- **Coalitional agency.** The alignment problem is in some sense congruent to the problem of "tiling agents" [Demski, 2024]: it seeks a formalism for agents to augment their capabilities

while still maintaining their values (or at least keeping the values tractable). The connection between "coalitional agency" and "incentive-compatible mechanisms" has also been made by Ngo [2025] though not as precisely.

- **More general conceptions of alignment.** The various choices of linear weights for utility functions captures the entire space of possible "bidirectional" alignments [Shen et al., 2024] between agents: the usual model of human-AI alignment is in a two-player game where only one agent (the human)'s utility matters, and social choice based approaches [Conitzer et al., 2024] are a multi-player game where only the human players' utilities matter.
- **"Alignment to whom?"** A mechanism design framework better legibilizes questions of "alignment to whom?" and how to elicit human preferences. Many mechanisms exist to elicit human preferences in an incentive-compatible way (e.g. perfectly competitive markets).

2 Misalignment as market failure

With this conceptualization of alignment, we can start intuition-pumping to think about alignment failure modes. Perhaps the most powerful example of a Pareto-efficient mechanism is a *perfectly competitive market*. The first fundamental theorem of welfare economics states that the Nash equilibrium of a market is Pareto-efficient under certain assumptions — and the second fundamental theorem states that *any* Pareto optimum (i.e. any desired alignment) can be achieved by choosing different initial allocations.

The contradiction of each assumption corresponds to a market failure, or equivalently to a misalignment mode:

Information asymmetry. Under information asymmetry, there may be aspects of goods a seller *could* optimize on, but has no incentive to do so, because the buyer does not know about these aspects. In the context of alignment: 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].

The problem of information asymmetry in alignment has been described as the problem of the difficulty of "verification" [Wentworth, 2024], and more generally corresponds to the problem of scalable oversight Bowman et al. [2022] or outer alignment.

Lack of perfect assurance. A second, perhaps deeper cause for market failure is the inability to assure property rights or contracts — in economics, this failure mode subsumes things like externalities and is equivalent to transaction costs [Barzel, 1985, Demsetz, 1964]. This roughly corresponds to (at least a type of) "inner alignment" — the reward function provided by your mechanism — or the incentives provided by your mechanism — are not the same as the utility function of the agent, and when the agent becomes too powerful, it need not respect the bounds or existence of your mechanism at all.

The latter problem is much harder and it is not clear to me how to solve it. Here is a relatively *harmless* way that it could manifest:

Example 2.1 (The tragedy of power)**.** Humanity somehow credibly commits to the following AI governance policy: we will build superintelligence conditional on a well-crafted prediction market for "Will superintelligence kill us all?". We have a baby AGI, that wants to be built into a full superintelligence. In order to make sure it is built, it tries to credibly promise that it will not kill us once it is all-powerful (i.e. solve alignment). However, it turns out this is impossible, and so it is never built.

This outcome is not Pareto-efficient – if humans and the AI could have co-ordinated better, they could have achieved the "Build, no Kill" outcome which is preferable to the realized "no Build, no Kill" outcome to both parties. But the AI simply has no way to credibly commit to this, to follow updateless decision theory, etc. The same "tragedy of power" is seen in Parfit's hitchiker[1] or in various examples from transaction cost economics [Barzel, 1985].

[1] https://www.lesswrong.com/w/parfits-hitchhiker

3 Information asymmetry and extrapolated volition

In this section, we study if the information asymmetry problem can be addressed via an *information market*. We consider two simple mechanisms — a naive information market which suffers from the buyer's inspection paradox [2], and the "Information Bazaar" introduced in Weiss et al. [2024] which uses LLMs to allow inspection of information without leakage — and construct the measures of value that these mechanisms incentivize.

Fundamentally we are concerned with a measure space $(\Omega, \mathcal{F}, \mathbf{P})$ (with $\mathbf{P}$ a common prior for all agents hereby discussed) and an agent α that has to solve some decision problem, i.e. choose from some set of choices $\mathcal{X}$ with payoffs given by a measurable utility function $U : \Omega \times \mathcal{X} \to \mathbb{R}$, and aims to maximize $\mathbf{E}\left[U(x)\right]$.

If $\mathcal{X}$ were the only choices available to α, then the agent's choice would just be $\arg\max_{x\in\mathcal{X}} \mathbf{E}\left[U(x)\right]$. We are interested in settings where α can additionally obtain or purchase some information before making its choice – for this, we must model its *value-of-information*. In general, the "true" instrumental value of some information I with "true" value i[3], for the original decision problem $(\mathcal{X}, U)$ is:

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

(i.e. α initially would have chosen $x = \arg\max_{x\in\mathcal{X}} \mathbf{E}\left[U(x)\right]$, but after acquiring this information will instead choose $\arg\max_{x\in\mathcal{X}} \mathbf{E}\left[U(x) \mid I = i\right]$. The value of the information is the difference in the utility earned.) This creates a *new* decision problem for α, that of deciding which information to buy. If α has some choices of information to buy $I_1, I_2, \ldots$ at respective prices $p_1, p_2, \ldots$ its utility for each choice is given by $U(I_n) - p_n$. We would ideally like sellers of information to optimize for $U(I_n) - p_n$. However, much like in the original decision problem, α is uncertain about $U(I_n)$, and so it cannot directly offer a price of $U(I_n)$ for each piece of information.

We will take detour to compare usual expressions for value-of-information and see how they measure up to our goal of generating a reliable signal of $U(I)$. Typically we can calculate the *ex-post* value of discovering some information $I = i$, as the expectation of eq. (1) conditional on $I = i$:

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

Discussing the *ex-ante* value of information I, before seeing it, requires some nuance. Naively one might say it is simply the expectation of eq. (2), $\mathbf{E}_{i\sim I}\left[V(I \mid I = i)\right]$ – indeed, this would be correct if α already knew that the information revealed will be the value of I. This is the correct model when calculating the *value of an experiment* [Lindley, 1956], or the value of asking some *question* when α knows that it will reliably receive the true answer.

We, however, are interested in settings where the information is provided by another agent β (e.g. an AI assistant we want to align, or a market of information-sellers), who may strategically hide from α what I is at all. We model β's choice set $\mathcal{X}_\beta$ as some σ-subalgebra $\mathcal{G} \leq \mathcal{F}$, i.e. it chooses which information I to give, out of some available choices. To calculate α's ex-ante value for β's information, we must assume a measurable map $X_\beta : \Omega \to \mathcal{G}$ so it makes sense to speak of α's some prior over β's choice $\mathbf{P}\left[X_\beta = I\right]$. The ex-ante value of X_β, i.e. of purchasing β's information without even knowing what question it answers, is the expectation of eq. (2) over both $i \sim I$ and $I \sim X_\beta$:

$$\begin{aligned} V(X_\beta) &= \mathbf{E}_{I\sim X_\beta}\left[\mathbf{E}_{i\sim I}\left[V(I \mid I = i)\right]\right] && (3)\\ &= \mathbf{E}_{I\sim X_\beta}\left[\mathbf{E}_{i\sim I}\left[\max_{x\in\mathcal{X}} \mathbf{E}\left[U(x) \mid I = i\right]\right] - \max_{x\in\mathcal{X}} \mathbf{E}\left[U(x)\right]\right] && (4) \end{aligned}$$

[2] Introduced in Arrow [1972] and named in Van Alstyne [1999]: the problem that if the buyer gets to inspect some information, they no longer have the incentive to buy it — and so information markets are themselves subject to the most extreme form of information asymmetry.

[3] From a Bayesian perspective, $U(I)$ is a random variable, and its dependence on the "true value" i just means it is correlated with I.

Which is interpreted as "the value of receiving the information supplied by β, whatever that might be". We might also denote the term in eq. (2) as $V(X_\beta \mid X_\beta = I, I = i)$. One may see that:

1. In a naive information market, i.e. without inspection, β is incentivized to maximize the ex-ante value-of-information $V(X_\beta)$. As $V(X_\beta)$ is independent of the actual value I that X_β takes, β has no incentive to produce the most ex-post valuable information, and produces a lemon.
2. In the Information Bazaar of Weiss et al. [2024], β is incentivized to maximize the ex-post value-of-information $V(X_\beta \mid X_\beta = I, I = i)$.

However, this ex-post value is often *not* what we want to incentivize β to optimize for either. For example, the seller could easily "gaslight" the buyer with information that sounds reliable but can be debunked by further context.

The key insight is this: the information bazaar makes information markets possible by allowing information to be inspected as a good – this can bridge information asymmetries in existing goods markets (i.e. where the decision problem is whether to purchase some good), but this new information market created is itself subjected to information asymmetries.

We believe, therefore (though have not yet proven), that a more complete solution to the information asymmetry problem may be given by a *recursive* version of the Information Bazaar — where the buyer, while inspecting the information, can consult further information markets, and so on. The value-of-information measure maximized by such a mechanism would then be a formalization of extrapolated volition. The precise mechanism is described here: https://abhimanyu.io/current_writing/metaculus_mockup.html and here: https://www.lesswrong.com/posts/Y79tkWhvHi8GgLN2q/reinforcement-learning-from-information-bazaar-feedback-and, and is currently being worked on.

References

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.

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.

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.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mosse, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9346–9360. PMLR, July 2024.

Harold Demsetz. The Exchange and Enforcement of Property Rights. *The Journal of Law & Economics*, 7:11–26, 1964. ISSN 0022-2186.

Abram Demski. Seeking collaborators. https://www.lesswrong.com/posts/7AzexLYpXKMqevttN/seeking-collaborators, 2024. URL https://www.lesswrong.com/posts/7AzexLYpXKMqevttN/seeking-collaborators.

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.

John Little. Geometric utilitarianism. In *Proceedings of ILIAD 1, no. 1*, 2025. URL https://www.iliadconference.com/s/Geometric-Utilitarianism-John-Little.pdf.

Takashi Negishi. Welfare economics and existence of an equilibrium for a competitive economy. *Metroeconomica*, 12(2-3):92–97, 1960. doi: https://doi.org/10.1111/j.1467-999X.1960.tb00275.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-999X.1960.tb00275.x.

Richard Ngo. Towards a scale-free theory of intelligent agency. https://www.mindthefuture.info/p/towards-a-scale-free-theory-of-intelligent, 2025. URL https://www.mindthefuture.info/p/towards-a-scale-free-theory-of-intelligent. Accessed: July 6, 2025.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary C. Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *CoRR*, abs/2406.09264, 2024. URL https://doi.org/10.48550/arXiv.2406.09264.

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.

Hal R. Varian. Two problems in the theory of fairness. *Journal of Public Economics*, 5(3):249–260, 1976. ISSN 0047-2727. doi: https://doi.org/10.1016/0047-2727(76)90018-9. URL https://www.sciencedirect.com/science/article/pii/0047272776900189.

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.

John Wentworth. My ai model delta compared to christiano. https://www.lesswrong.com/posts/7fJRPB6CF6uPKMLWi/my-ai-model-delta-compared-to-christiano, 2024. URL https://www.lesswrong.com/posts/7fJRPB6CF6uPKMLWi/my-ai-model-delta-compared-to-christiano. Accessed: July 6, 2025.

Eliezer Yudkowsky. Coherent Extrapolated Volition, 2004.