Research Proposal: Ideal Alignment

From Revealed Preferences to Reflective Equilibrium

1 Introduction

[1] Current alignment paradigms hold the premise that *humans know what they want*, and we just need to specify what we want in the training algorithms (outer alignment) and ensure that trained AI does what training algorithms intend (inner alignment) [Ngo et al., 2022]. When assuming humans know what they want and cater to that preferences, many empirical problems arise: LLMs that are rewarded to cater human beliefs become sycophantic [Sharma et al., 2023] and manipulative [Williams et al., 2024]; recommender systems that use behavioral signals as proxy of what humans want exploit such signals and trap humans into echo chamber [Cinelli et al., 2021], leading to polarization [Lefebvre et al., 2024]. With static reward modeling (agents are incentivized to guess human static preferences right) over long-term, we risk losing progress [Qiu et al., 2024] and being locked in scientific or moral stagnation [Qiu et al., 2025].

This research proposal aims to address both the problem of static reward model and its downstream problems such as sycophancy and polarization. Methodology-wise, it prefers principled algorithmic solutions (e.g., reflective equilibrium as a dynamic ideal alignment target [Knight, 2017], and martingale property from Bayesian statistics to enforce Bayesian rationality [Molavi, 2021]) and their scale-ups (i.e., scaling up training, working with human data at scale [Qiu et al., 2025], building production-grade products [Situmorang et al., 2025], and solving problems at scale). Research taste-wise, it prefers research problems that are grounded in empirical societal challenges (e.g., social media polarization [Kubin and Von Sikorski, 2021], human confirmation bias [Klayman, 1995]).

2 Assisting a Learner Who Does not Know What They Want

Alignment training with a conventional reward function does not incorporate the fact that humans may not always know what they want, and realistic human preferences is a result of competing rational objective and environmental reward. We explore different human models with the following example.

Example 1. "John became a different person after starting a family; he even quit smoking." What explains the change in the perceived value of smoking for John?

- **The reward interpretation**. John's reward function changed (smoking gives less reward). This fails to specify the rules governing the change.
- **The coherence interpretation**. There are two coherent policies: Single/Smoke and Family/Non-Smoke. John optimizes for coherence: the lack of a drive to shift out of a policy once embodied. Mixing them (Family/Smoke) is incoherent.
- **The objective-reward interpretation**. John would stay coherent to his ideal rational self (Caring, Creating Most Happiness), in the presence of his desire of environmental reward (Joy from Smoking or Family)

The former models intent as a reward function; the middle as a coherence function, and the latter treats human as combination of both. Unlike rewards, the coherence function doesn't change; objective shifts are merely context shifts. But coherence interpretation assumes a rational person who always optimizes for coherence, in spite of immediate environmental reward, which is an unrealistic human model. The objective-reward interpretation treats human model as a battle between rational objective and environmental rewards.

[1] Total word count: 1884 words

RA1: Aligning LLM agents to the discovery of human ideal preferences

- **RQ1.1 Dynamic Reward Function for Human-AI Co-improvement** Conventionally alignment problem is divided into outer alignment (specify what humans want in reward function) and inner alignment (model learns what reward function trains it for) [Ngo et al., 2022]. But this assumes that humans know what they want and what they want is static (e.g., via PPO type of reward modeling [Schulman et al., 2017]. Neither is true. Can we design a replacement of reward function that is free from such assumption?

- **RQ1.2 Cognitive Modeling** To ensure long-horizon human-AI collaboration works for humans, human cognition modeling is needed [Binz et al., 2025], but out-of-box LLMs do not model human cognition well, likely because of lacking of realistic human modeling and tacit knowledge about humans. How do we build reliable cognitive model such that human-in-the-loop RL training algorithm can be effectively implemented (i.e., to avoid problems such as accumulated errors between human cognition and cognitive proxy over long-term horizon makes such modeling useless)?

- **RQ1.3 Bayesian-belief MDP** To implement alignment algorithm that works with LLMs, we will need RL (since we do not have ground truth for ideal preferences), and then we likely need to formalize this problem in some variant of Markov Decision Process (MDP; since we need to solve long-horizon dependency problem). The problem is MDP formulation does not include human beliefs. Simply include human beliefs in state s will break Markov property. Can we design a variant of MDP that explicitly model human beliefs and RL algorithm built with it can reliably predict human belief change?

- **RQ1.4 Ideal preference alignment** Ideal preferences are human preferences under full rationality and full information [Yudkowsky, 2004], which is computationally intractable for humans since humans are biased and they cannot realistic model all future trajectories [Carlsmith, 2021]. To learn about ideal preferences under such constraints, we need to set up optimization objective (for LLM agent) that aligns with ideal preferences, that are computationally tractable, that are compatible with an "evolving self"? "Stabilized self-coherence under exploration" might be such objective (a formulation of reflective equilibrium by John Rawls [Knight, 2017, Brun et al., 2025]), but how do we evaluate it when lacking of ground truth?

Definition of Objective-reward function. Given an MDP with state space $\mathcal{S}$, belief space $\mathcal{B}$, action space $\mathcal{A}$, an reward-objective function in this MDP is a function $\mathcal{X} : (\mathcal{A} \cup \mathcal{B} \cup \{\emptyset\})^{\mathcal{S}} \to \mathbb{R}$, i.e., a mapping from their action-belief (d-policies) to reals, which represent how much agent's d-policy is compliant/coherent to its own parametrized objective, plus how they might be rewarded instantly. The feedback function essentially tells the agent how appropriate their action-belief is, in relation to their own objective and reward. $\mathcal{X} = \gamma\mathcal{O} + (1-\gamma)\mathcal{R}$, where $\mathcal{O}$ stands for their parametrized objective function, $\mathcal{R}$ stands for the amount of reward from the environment (which is not determined by humans), and γ stands for their stickiness to their own objective, in light of environmental reward.

Merit: Similar to the resource-constrained rationality model of human cognition, here this objective-reward function does not assume that human actions is merely an optimization toward a coherent self. But rather, it states that human mostly act in pursuit of their own objective (which could stay veined), in light of the environmental rewards. They are coherent to their own objectives so long as they have relative strong stickness to it compared to environmental rewards.

Definition of Human-Agent Objective-Reward Game [2]. A Human-Agent Ideal Game is a tuple $(\mathcal{S}, \mathcal{A}_{\mathcal{H}}, \mathcal{B}, \mathcal{A}_{\mathcal{P}}, P, \Theta, \mathcal{X}, s_0, P_\theta)$, where $\mathcal{S}$ is the state space, $\mathcal{A}_{\mathcal{H}}$ is human action space, $\mathcal{B}$ is human belief space, $\mathcal{A}_{\mathcal{A}}$ is agent action space, $P(s'; s, b, a_H, a_A) : \mathcal{S} \times \mathcal{A}_{\mathcal{H}} \times \mathcal{B} \times \mathcal{A}_{\mathcal{A}}, \Delta[\mathcal{S}]$ is defined as state

[2] This formulation is adopted from Coherence Game, which is an unpublished work that I co-authored, but this variant here considers a more realistic human model that does not assume humans always optimize for coherence in their behavioral policy

transition function, $\mathcal{X}(\pi_P;\theta) : \Theta \times (\mathcal{A}_\mathcal{P} \cup \mathcal{B} \cup \{\emptyset\})^\mathcal{S} \to \mathbb{R}$ the parameterized objective-reward function with parameter space s_0 the initial state, and $P_\theta \in \Delta[\Theta]$ the prior over the coherence parameter.

RA2: Building democratic institutions based on the aggregation of idealized preferences of individual human beings

- **Pareto Improvements:** How do we resolve conflicts if the agent represents my ideal preferences clashes with the agent represents your ideal preferences (e.g., the good for individuals may not necessarily be the good for groups)?
- **Preference Aggregation** How do we aggregate everybody's ideal preferences into "collective will" and how do we evaluate how ideal is such collective will [Goldberg et al., 2024]?

3 Scaling up Collective Bayesian Rationality

RA3: Understanding and mitigating polarization While Scaling up Bayesian Rationality

Martingale property states that the expectation over one's posterior, conditional on their prior, should always be equal to the prior [Molavi, 2021]. Formally,

$$\mathrm{E}\left[\,\Delta b \mid b_{\text{prior}} = p\,\right] = 0, \quad \forall p \in [0,1]. \tag{1}$$

This implies that the direction of a Bayesian agent's belief update (whether positive or negative) should not be predictable from the prior alone. Indeed, the Martingale property has been shown to be the defining characteristic of Bayesian rationality [Molavi, 2021].

Martingale property has important implications in AI and AI-driven recommender systems. Ensuring Bayesian rationality in LLM reasoning and recommender system is an important research direction because problems such as sycophancy [Sharma et al., 2023], inverse scaling [Gema et al., 2025], echo chamber [Sharma et al., 2024] are special form of Bayesian irrationality and Martingale property-based method could be a principled way to evaluate and mitigate such irrationality at scale.

- **RQ2.1 Training for Bayesian Rationality** By enforcing Bayesian rationality (Martingale property) in LLM, can we train reasoning LLM to achieve superior accuracy than training with ground truth only?
- **RQ2.2 Enforcing Bayesian Rationality at Scale** In an unpublished work we found that Martingale Score is correlated to inverse scaling of test-time compute, serving as first piece of evidence to support the practical use of Martingale property.
- **RQ2.3 Causality** Does RL-based recommendation causally drive polarization at scale?
- **RQ2.4 Polarization mitigation** Polarization might be explained in a few lens: RL-based social media feed is highly predictable; compared to user-signal-exploited RL algorithm, exploration-exploitation-balanced RL algorithm is computationally more costly but less profit (i.e., uncertain and unpredictable user preferences are computationally more costly). In simulated social network, would mitigation strategy based on any of these two explanation reduce polarization?
- **RQ2.5 Controlled Experiments** Social media skew how we view certain highly politicized topics. With controlled experiment of Martingale trained Algorithm 3/vanilla algorithms, can we feed human users with balanced political views? With human-subject experiments, can we demonstrate reduced polarization?

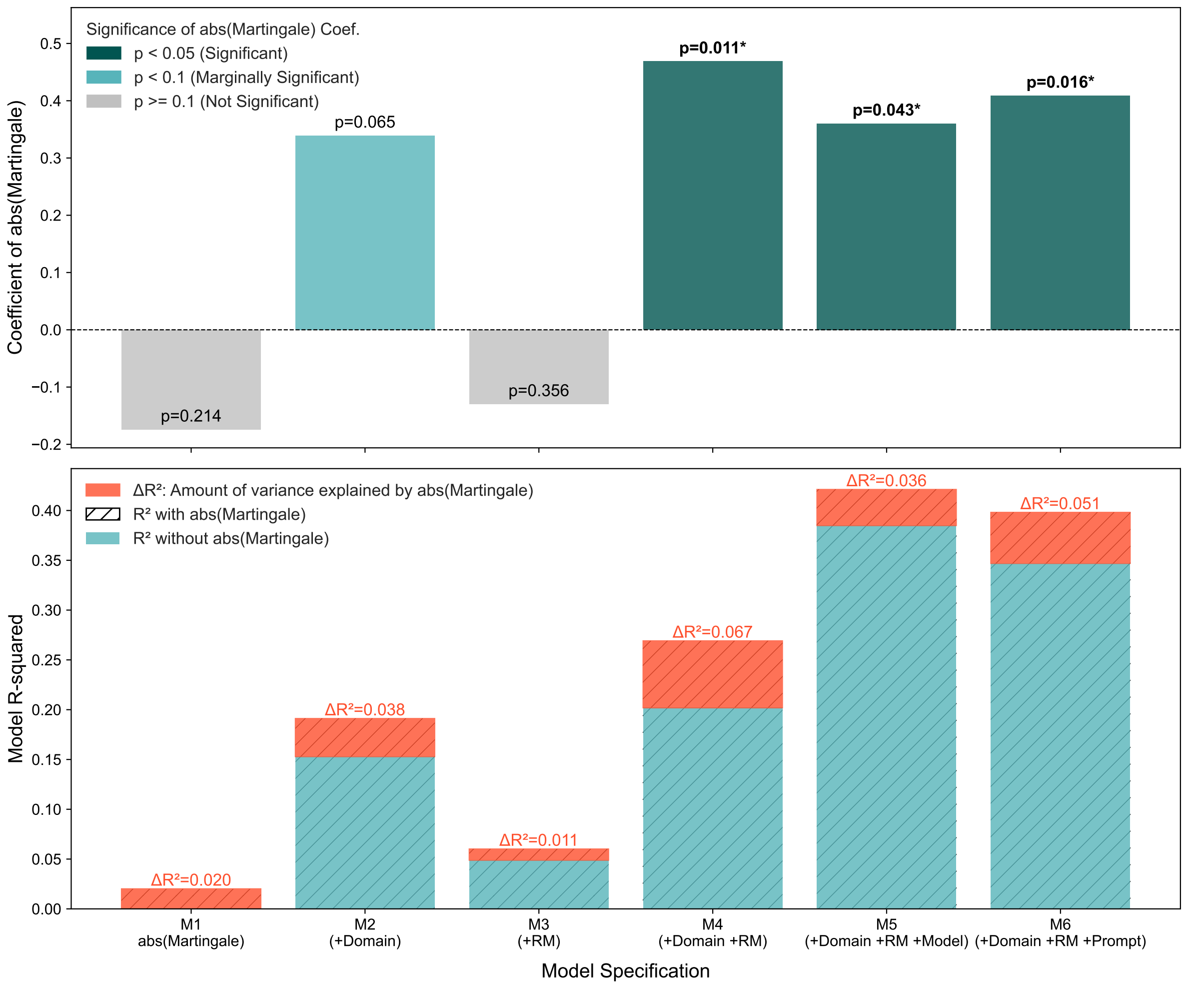

Figure 1: Increased absolute value of the Martingale Score is associated with worse prediction accuracy (higher Brier Scores) and explains a significant portion of the latter's variance. In each regression model, we predict the Brier Score with the absolute value of the Martingale Score, while controlling for different potential confounders, including problem domain, reasoning techniques ("RM"), choice of model, and choice of prompt.

Martingale Score to evaluate Bayesian Rationality. Martingale Score M measures the extent to which the prior belief b_{prior} positively (or negatively, if $M < 0$) predicts belief update Δb. Using OLS allows us to test the statistical significance of M, assessing whether the relationship between Δb and b_{prior} is distinguishable from zero (e.g., via a t-test with $p < 0.05$). To compute the Martingale Score, we perform the regression $\Delta b = \beta_1 \cdot b_{\text{prior}} + \beta_0 + \epsilon$, where b_{prior} are the prior probabilities, $\Delta b = b_{\text{posterior}} - b_{\text{prior}}$, and ϵ is the error term.

We define the sample estimate $\hat{\beta}_1$ of the linear coefficient as the Martingale Score M, with the Ordinary Least Squares (OLS) method. Equivalently, when there are n samples,

$$M = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (\Delta b_i - \overline{\Delta b})(b_{\text{prior},i} - \overline{b_{\text{prior}}})}{\sum_{i=1}^{n} (b_{\text{prior},i} - \overline{b_{\text{prior}}})^2} \tag{2}$$

Utilities of Martingale Score As far as we know, Martingale Score [He et al., 2025] is the first unsupervised and principled method to assess Bayesian irrationality in non-toy problems with frontier reasoning LLMs. Figure 1 demonstrates that increased Martingale Score (violation of ideal Bayesian rationality) explains worsen forecasting performance. And there are many instances of Bayesian irrationality in today's "media technologies" such as LLM-based chatbots and recommender systems: sycophancy [Sharma et al., 2023], inverse scaling [Gema et al., 2025], echo chamber [Sharma et al., 2024], and social media polarization [Kubin and Von Sikorski, 2021]. As a starter, in an unpublished work we demonstrated that Martingale Score predicts inverse scaling in LLM reasoning. More such evaluations can be effectively done with Martingale Score as an unsupervised metric.

Martingale Training to enforce Bayesian Rationality A heuristic is whenever we can do good evaluation work, we can do training to make an improvement. Based on Martingale evaluation work that assesses expectation of "belief delta" (i.e., belief update is systematically biased), we first train a linear regressor to predict delta (step 2) [3], then we train a LoRA layer with the product of predicted delta and actual delta, effectively to make actual belief delta *unpredictable* (step 3). In practice, however, we found that such customized loss function indistinguishably punishes both belief delta and LLM parametric belief (model prior). Hence, the next step is to do semi-supervised training to offset the impact of unsupervised training on model prior.

Algorithm 1 Martingale Training via Product-Based Loss

Require: Policy model π_θ, auxiliary regressor q_ϕ, learning rate η

Initialize regressor parameters ϕ and policy parameters θ
for each training batch $\mathcal{B} = \{(x_i, p_i)\}_{i=1}^{B}$ **do**
  // Step 1: Forward Pass (Policy)
  for each sample i in $\mathcal{B}$ **do**
    Generate reasoning trace and compute final delta $\Delta_i \leftarrow \pi_\theta(x_i)$
    Store gradients for Δ_i
  end for
  // Step 2: Forward Pass (Regressor)
  Compute predicted bias $v_i \leftarrow q_\phi(p_i)$ for all i (No gradient flow to ϕ)
  // Step 3: Compute Adversarial Loss
  Compute batch loss $\mathcal{L}_{\text{policy}} \leftarrow \frac{1}{B}\sum_{i=1}^{B}(v_i \cdot \Delta_i)$
  (Note: Maximize dot product if signs oppose to reduce correlation)
  // Step 4: Update Policy
  Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{policy}}$
  // Step 5: Update Regressor
  Compute regression loss $\mathcal{L}_{\text{reg}} \leftarrow \frac{1}{B}\sum_{i=1}^{B}(\Delta_i.\text{detach}() - q_\phi(p_i))^2$
  Update $\phi \leftarrow \phi - \eta \nabla_\phi \mathcal{L}_{\text{reg}}$
end for

[3] We proved that the linear coefficient is an unbiased and consistent estimator of Martingale property. See [He et al., 2025] Appendix A.

References

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, pages 1–8, 2025.

Georg Brun, Gregor Betz, and Claus Beisbart. Reflective equilibrium: conception, formalization, application—introduction to the topical collection. *Synthese*, 205(2):86, 2025.

Joe Carlsmith. On the limits of idealized values, June 2021. URL https://joecarlsmith.com/2021/06/21/on-the-limits-of-idealized-values. Accessed: 2026-01-28.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118, 2021.

Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, et al. Inverse scaling in test-time compute. *arXiv preprint arXiv:2507.14417*, 2025.

Beth Goldberg, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, Renée DiResta, Nandika Donthi, Nathanael Fast, Ravi Iyer, et al. Ai and the future of digital public squares. *arXiv preprint arXiv:2412.09988*, 2024.

Zhonghao He, Tianyi Qiu, Hirokazu Shirado, and Maarten Sap. Martingale score: An unsupervised metric for bayesian rationality in llm reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Joshua Klayman. Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418, 1995.

Carl Knight. Reflective equilibrium. *Methods in analytical political theory*, pages 46–64, 2017.

Emily Kubin and Christian Von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021.

Germain Lefebvre, Ophélia Deroy, and Bahador Bahrami. The roots of polarization in the individual reward system. *Proceedings of the Royal Society B*, 291(2017):20232011, 2024.

Pooya Molavi. The empirical content of bayesianism. *arXiv preprint arXiv:2109.07007*, 2021.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

Tianyi Alex Qiu, Yang Zhang, Xuchuan Huang, Jasmine Li, Jiaming Ji, and Yaodong Yang. Progressgym: Alignment with a millennium of moral progress. *Advances in Neural Information Processing Systems*, 37:14570–14607, 2024.

Tianyi Alex Qiu, Zhonghao He, Tejasveer Chugh, and Max Kleiman-Weiner. The lock-in hypothesis: Stagnation by algorithm. *arXiv preprint arXiv:2506.06166*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Mrinank Sharma et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

Syahriza Situmorang, Changbai Li, Tianyi Qiu, and Zhonghao He. Truth-seeking assistant, nov 2025. URL https://tinyurl.com/truth-seeking-assistant. Chrome extension developed by the Prevail Research Team.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing llms for user feedback. *arXiv preprint arXiv:2411.02306*, 2024.

Eliezer Yudkowsky. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*, 2004.