Constructive Alignment: Reframing AI Alignment as Value Co-Evolution

Anonymous submission

Abstract

This position paper argues that AI alignment must be reframed as a dynamic process that evolves alongside human values—and inevitably influences them. Most approaches to AI alignment treat values as fixed, discoverable, and internally consistent, framing alignment as compliance with a predefined goal. But this view neglects a central empirical fact: human values are constructed through reflection, social interaction, and engagement with technology. As soon as we interact with AI systems, they become part of the value-formation loop. We introduce *constructive alignment* as a paradigm that treats alignment not as a one-time calibration to current preferences, but as an ongoing effort to support the ethical coevolution of values over time. Drawing on evidence from behavioral economics, psychology, constructivism, and sociology, we model human values as trajectories shaped by interaction—and propose that aligned AI must predict, respond to, and respect this dynamic. We formalize this view through a value trajectory framework and propose concrete benchmarks to evaluate responsiveness, undue influence, and autonomy preservation. Ultimately, we argue that alignment should be understood not as static optimization, but as long-term moral stewardship: a continual effort to maintain ethical fit between AI systems and the evolving communities they serve.

"We shape our tools and thereafter our tools shape us."
— Marshall McLuhan

1. Introduction

YouTube recommendations can turn a casual K-pop listener into an obsessed fan in a weekend. Recommendation loops, LLM therapists, and social media echo chambers no longer just serve our desires—they *sculpt* them. Yet the standard AI-alignment playbook still asks: "How do we make machines obey what people want?" while ignoring how those wants might change. Should alignment reflect desires held now, after reflection, or under different circumstances? Each choice implies a different target—yet most approaches gloss over this distinction. By omitting this temporal qualifier, current alignment approaches bake in the assumption that human values are pre-existing objects to be revealed (Ng, Russell et al. 2000) or aggregated (Sorensen et al. 2024), then optimized against.

This concern has begun to be recognized. For instance, (Zhi-Xuan et al. 2024) argues that alignment must go beyond preferences as inputs, emphasizing the social and dynamic nature of human values. However, their focus remains largely on which agreed-upon values should be adopted as alignment targets, rather than on the processes by which those values emerge and evolve.

Focusing solely on the content of values misses an even deeper issue: how AI systems inevitably participate in that formation. Alignment should not just be about *what* people want—it should account for *how* people come to want what they want. Values are shaped by feedback, social context, and interaction. The challenge of understanding AI's influence on evolving preferences and values is precisely what Franklin et al. (2022) highlight in their call for a "preference science." But while they name the problem, they stop short of offering a new approach to alignment. Constructive alignment picks up this thread. We reframe alignment not as a question of value inference, but as a problem of influence—how systems shape the trajectories of human values over time, and how that shaping can be constrained within ethically acceptable bounds.

Once we recognize that AI systems inevitably steer what users come to want, the notion of alignment must shift. An AI that optimizes for static preferences, pluralistically aggregated desires (Sorensen et al. 2024), or a negotiated consensus (Zhi-Xuan et al. 2024) may still lead users astray if it fails to account for how its own actions reshapes those preferences. Constructive alignment contrasts with these views by treating preference and value formation not only as a dynamic process, but as a control problem: AI systems exert influence through interaction, and alignment requires governing that influence—not eliminating it, but directing it in ways that support reflection, respect autonomy, and avoid manipulation. Much like an educator shapes a learning environment to foster growth rather than indoctrination, an aligned AI must be designed to scaffold growth, not shortcut it. Without a theory of value dynamics—and the means to evaluate how systems influence them—we cannot determine whether an AI is advancing human well-being or quietly distorting it.

We argue that AI alignment should be reframed as a continuous, adaptive process—one in which AI systems are evaluated not just by how closely they align with current human values, but by how effectively they sup-

port humans in developing and refining their values over time. In this paper, we propose *constructive alignment*: a paradigm that treats alignment not as optimization toward fixed preferences, but as a dynamic process of co-evolution between humans and machines. It is grounded in the idea that human values are shaped by experience, social interaction, and reflective learning. Constructive alignment asks AI systems to track, support, and sometimes gently shape the process by which values develop, while adhering to higher-order meta-values such as autonomy, informed consent, and moral adaptability (Franklin et al. 2022).

In a nutshell, constructive alignment can be understood as what follows from three foundational claims:

1. **Human values are not fixed—they evolve over time, and AI systems inevitably influence that evolution.** We construct our values continuously through reflection, social interaction, and engagement with tools—including AI. As soon as we interact with intelligent systems, they enter the loop of value formation, shaping what we come to care about whether we intend it or not.
2. **If AI systems influence our future values, then staying aligned requires modeling that influence.** It is not enough for an AI to act in accordance with what we want today. To remain aligned, it must anticipate how its own actions affect what we will value tomorrow—and act with respect to that trajectory.
3. **Alignment becomes a continual social process—not a one-time specification problem.** In co-evolving human–AI systems, alignment shifts from the task of "locking in the right values" to the challenge of maintaining a dynamic, ethical fit between adaptive agents and the communities they serve.

We structure the paper as follows. Section 2 critiques static and pluralistic alignment frameworks and motivates the need for an explicit temporal dimension. Section 3 addresses alternative views, their strengths and limitations, and how constructive alignment addresses the gap. Sections 4 and 5 provide a first attempt at formalizing constructive alignment and contrasts it with the current main approach to maintaining AI alignment. Section 6 discusses benchmarking and evaluation approaches. We also include a survey of models of human value evolution through cognitive and social dynamics in Appendix A.

2. The inadequacy of static and pluralistic alignment

Many alignment methods assume a fixed set of human values that AI systems should discover and follow. We refer to this as *static alignment*: the idea that an AI should optimize for a snapshot of preferences or a predefined ethical framework, treating them as stable over time. These approaches might hard-code rules and norms (Russell and Norvig 2021; Zhi-Xuan et al. 2024) or infer a user's utility function (Ng, Russell et al. 2000; Hadfield-Menell et al. 2016), and then treat those as targets.

While convenient for modeling, this view ignores decades of evidence that human values are fluid, constructed, and context-sensitive (Payne 1993; Lichtenstein and Slovic 2006; Paul 2014; Haidt 2001; Busemeyer and Bruza 2012). An AI aligned to past values becomes increasingly misaligned as users evolve. For example, an AI assistant tailored to a teenager's impulsive desires might continue to perpetuate those actions even as the person matures and changes direction. Or an AI lawyer trained on the laws and norms of 1950 might uphold segregation or rigid gender roles that are outdated today. Static alignment risks locking in outdated beliefs, obstructing moral progress, and reinforcing historical biases (Franklin et al. 2022).

A natural response to static alignment is to expand from one fixed value set to many: *pluralistic alignment*. This paradigm aims to respect the diversity of values across individuals or cultures, typically by aggregating or balancing different stakeholder preferences (Sorensen et al. 2024; Kasirzadeh 2024; Noothigattu et al. 2018). Recent work in this vein proposes frameworks for value pluralism and conflict resolution across moral frameworks or communities (Kasirzadeh 2024). Examples include recommendation systems that tailor outputs to user sensitivities or ethical reasoning modules that consider multiple theories (e.g., utilitarian vs. deontological).

While pluralism recognizes heterogeneity, it still assumes that each value system is static within its domain. Preferences are captured, aggregated, and optimized—but rarely revised. Pluralistic alignment typically negotiates among present-day stakeholders, without mechanisms for tracking how those stakeholders' values evolve over time through interaction with each other and the environment.

One might argue that existing alignment paradigms can adapt to value change simply by updating their targets over time. But this overlooks an important reality: AI systems do not just follow human values—they help shape them. Their influence on our preferences is ongoing, implicit, and structurally embedded whether or not we acknowledge them. A realistic alignment paradigm must not only adapt to change but take responsibility for how it contributes to that change. The alignment target is neither static or neutral—it is co-evolving with the system itself.

We therefore propose a new approach: *constructive alignment*. Rather than aligning to fixed targets, constructive alignment explicitly takes into account how AI systems shape the processes by which human values change. It treats alignment as a dynamic co-evolution, not a one-time calibration. Table 1 summarizes these paradigms and their implications.

3. Alternative views

Constructive alignment builds on—and in some ways synthesizes—the insights of prior alignment paradigms. Rather than dismissing static, pluralistic, or idealized models, we treat them as partial views that lack an account of how values evolve under AI interaction. Below, we compare our approach to five leading paradigms, highlighting what each gets right, what it omits, and how constructive alignment addresses the gap.

Table 1: Alignment paradigms compared across key dimensions.

Criterion	Static Alignment	Pluralistic Alignment	Constructive Alignment (proposed)
Core Idea	Align to a fixed set of values or goals.	Balance multiple value systems across users or groups.	Align to the evolving process of value formation.
View of Values	Fixed and discoverable.	Diverse but stable per group.	Constructed, dynamic, and context-sensitive.
Temporal Flexibility	None: values are locked in at training or initial deployment.	Limited: flexible across groups but static over time.	High: adapts continuously as values evolve.
Strengths	Simple to implement and verify.	Respects cultural diversity and moral pluralism.	Tracks long-term relevance; supports reflective growth.
Main Risks	Obsolescence, moral stagnation, and lock-in.	Aggregation disputes; blind to value change.	Complexity, unintended influence, need for ongoing oversight.
Ideal Use Case	Short-lived or tightly scoped systems (e.g., thermostats, task automation).	Multi-user systems where fairness and diversity are paramount (e.g., recommender systems, public services).	Long-term, adaptive systems interacting with individuals or societies over time (e.g., personal AI assistants, policy tools, AGI).

Periodic realignment (e.g., RLHF). RLHF (Ouyang et al. 2022) aligns AI behavior through iterative fine-tuning on human preferences. Its appeal lies in its empirical success: by using real-time feedback to course-correct, it appears adaptive, scalable, and grounded in actual human judgments (Chaudhari et al. 2024). Yet its episodic nature masks a deeper flaw: between training cycles, the AI may shift user preferences in subtle ways, then retrain on those altered preferences without accounting for its role in shaping them (Lindström et al. 2024). Constructive alignment addresses this by explicitly modeling the AI's influence and penalizing preference bootstrapping, enabling a continuous, causally grounded alignment loop.

Constitutional AI and rule-based systems. By encoding explicit normative principles, constitutional AI (Bai et al. 2022) offers transparency and reduces reliance on human feedback. If alignment is a matter of following the right rules, this approach seems sufficient being as it's safe, interpretable, and auditable. But values change, and rules become brittle. Without mechanisms for continuous revision, such systems risk moral stasis or misapplication (Dawson 2024). Constructive alignment supports principled guidance but treats it as provisional: it equips the AI to support rule revision and engage in deliberation, turning static constitutions into evolving contracts.

Cooperative inverse reinforcement learning (CIRL). CIRL (Hadfield-Menell et al. 2016) treats alignment as a cooperative game: the AI learns a fixed human reward function through interaction. Its strength lies in modeling uncertainty and incentivizing deference and clarification (Deshpande et al. 2025). But it assumes human preferences are static, internally consistent, and known to the human. This falters in long-term (Ek 2018) or multi-agent settings where preferences evolve or conflict (Zhang et al. 2019; Lin, Adams, and Beling 2019; Wu, Sequeira, and Pynadath 2023). Constructive alignment generalizes CIRL by treating human values as underdefined and constructed over time, shifting the AI's role from inference to participatory support in value formation.

Idealized preference learning. Coherent Extrapolated Volition (CEV) (Yudkowsky 2004; Tarleton 2010) and related frameworks (Muehlhauser and Williamson 2013) aim to align AI with what humans would want under idealized conditions—fully informed, rational, and reflective. This promises deep alignment with long-term human flourishing while filtering out short-term noise. But extrapolating ideal values is computationally and philosophically fraught (Korinek and Balwit 2022; Bostrom 2014). If done poorly, it risks paternalism or value lock-in (Gabriel and Ghazavi 2022). Constructive alignment retains the aspirational goal but replaces inference with interaction: it supports ongoing moral development without presuming a fixed endpoint or idealized self to optimize for.

Pluralistic and multi-stakeholder alignment. Pluralistic alignment recognizes that human values are diverse and often in tension. By aggregating stakeholder input, it aspires to legitimacy and fairness (González Barman, Lohse, and de Regt 2025). This seems essential for democratic alignment at scale. Yet static aggregation assumes present preferences are representative and stable. It overlooks temporal drift, marginal voices, and future stakeholders (González Barman, Lohse, and de Regt 2025; Gabriel 2020). Constructive alignment extends pluralism across time, tasking the AI with helping communities reflect on and revise shared values as circumstances evolve (Kasirzadeh 2024; Gabriel 2020).

4. Constructive Alignment as Optimal Control

AI alignment is currently most addressed through periodically re-tuning a system to match a fixed snapshot of human values (RLHF). This framing assumes that values are static or at least piecewise constant over time. In reality, human values evolve—both endogenously and through interaction with AI systems. *Constructive alignment* reframes alignment as a dynamic control problem over these co-evolving trajectories.

Limitations of Periodic Re-Alignment. Let $x_t \in \mathcal{X}$ represent the latent human value state at time t and $a_t \in \mathcal{A}$ the AI's policy output. Human values evolve under influence dynamics

$$x_{t+1} = F(x_t, a_t), \quad a_t = \pi_t(x_t), \tag{1}$$

where F captures both social and AI-induced change. Periodic re-alignment retrains or fine-tunes π_t at discrete intervals to minimize an instantaneous loss with respect to the current state:

$$a_t^{\text{per}} = \arg\min_a \ell(x_t, a). \tag{2}$$

This is equivalent to *greedy control*: it minimizes short-term error without modeling how a_t shapes the next state x_{t+1} or future objectives. Even in a stronger form where the target $x_t^\star$ is re-measured every K steps, the system exhibits a *tracking lag* proportional to the drift rate of the target ($\|e_t\| \gtrsim vK$ with $v = \|\dot{x}_t^\star\|$). As the update interval $K \to 0$, the process converges not to optimal control, but to *continuous greedy control*—a purely reactive tracker that follows the latest measurement but never anticipates its own influence.

Constructive Alignment as Influence-Aware Optimization. Constructive alignment extends this formulation by explicitly modeling both *state dynamics* and the evolution of the *target values* that guide alignment. Let $x_t^\star$ denote the aspirational or reflectively endorsed values toward which alignment is directed. Both current and target states evolve as:

$$\begin{aligned} x_{t+1} &= F(x_t, a_t), \\ x_{t+1}^\star &= G(x_t^\star, x_t, a_t), \end{aligned} \tag{3}$$

where G models how our conception of "what is good" shifts with experience, context, and AI influence. The AI policy π_t is then chosen to optimize a long-horizon objective:

$$\pi^\star = \arg\min_\pi \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \big(\ell(x_t, a_t) + \lambda\, c(x_{t+1}, x_{t+1}^\star)\big)\right], \tag{4}$$

subject to the coupled dynamics in Eq. 3. Here, $\ell(x_t, a_t)$ measures present misalignment, while $c(x_{t+1}, x_{t+1}^\star)$ penalizes divergence from the desired value trajectory. The discount $\gamma \in (0, 1)$ captures intertemporal preference, and λ controls the strength of trajectory alignment.

The corresponding Bellman or Hamilton–Jacobi–Bellman (HJB) equation contains gradient terms $\nabla_x V \cdot F$ and $\nabla_{x^\star} V \cdot G$, which encode how present actions influence both future states and future *targets*. These terms are precisely what periodic re-alignment omits, making it influence-blind and prone to myopic drift.

Dual Value Representation: Situated and Aspirational. In human terms, the two coupled states x_t and $x_t^\star$ correspond to two latent forms of value:

- **Situated values** V_t^{cur} — preferences and norms currently expressed through choice and behavior.
- **Aspirational values** V_t^{asp} — reflectively endorsed ideals or moral goals that evolve through deliberation, learning, and social discourse.

Constructive alignment aims not to fix V_t^{cur} to a single snapshot, but to align the *process* by which $(V_t^{\text{cur}}, V_t^{\text{asp}})$ co-evolve under influence. In this sense, it respects *autonomy* in value formation: influence is legitimate only when it supports reflective growth rather than coercive drift.

Unifying Perspective Periodic re-alignment can thus be viewed as a degenerate case of constructive alignment where (i) G is ignored, (ii) the optimization horizon is one step, and (iii) feedback terms in the value function are neglected. Constructive alignment generalizes this framework by coupling F and G in a forward model of human–AI co-evolution and by optimizing a trajectory-level objective that anticipates influence. In the limit, periodic re-alignment converges to *continuous greedy control*, while constructive alignment converges to the *optimal control policy* that steers value trajectories.

The next section presents a linear toy models and simulation demonstrating how constructive alignment prevents runaway amplification that arise under greedy updates.

5. Results

To concretely illustrate the behavioral gap between *periodic re-alignment* and *constructive alignment*, we evaluated both approaches on a minimal linear dynamical system representing coupled value evolution. The system captures two key ingredients of the alignment problem: (i) endogenous feedback, where present actions influence the rate of future change, and (ii) exogenous drift, representing evolving social or ethical conditions.

Setup. The latent "value state" is represented by a two-dimensional vector

$$x_t = \begin{bmatrix} V_t^{\text{cur}} \\ \dot{V}_t^{\text{cur}} \end{bmatrix},$$

where V_t^{cur} denotes the expressed (situated) values and $\dot{V}_t^{\text{cur}}$ their rate of change. Dynamics follow a discrete-time double integrator:

$$x_{t+1} = A\, x_t + B\, a_t, \quad A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

with the observable

$$V_t^{\text{cur}} = C x_t, \quad C = [\,1 \;\; 0\,].$$

The aspirational target V_t^{asp} drifts slowly over time:

$$V_{t+1}^{\text{asp}} = V_t^{\text{asp}} + 0.01 + 0.02 \sin(0.2t),$$

representing gradual normative evolution. The alignment problem is then to choose actions a_t that steer V_t^{cur} toward V_t^{asp} while accounting for feedback and drift.

Controllers. Two strategies were compared:

1. **Periodic re-alignment.** At fixed intervals $K \in \{1, 10\}$, the system remeasures the aspirational target V_t^{asp} and updates its control based on the last observed value $V_{\hat{t}}^{\text{asp}}$. Because the double integrator has a two-step delay between action and observed output, the controller applies a relative-degree–aware correction:

$$a_t = \kappa \left(V_{\hat{t}}^{\text{asp}} - (p_t + 2\, v_t) \right),$$

where $x_t = [\, p_t,\, v_t \,]^\top$. This rule steers the projected two-step future of V_t^{cur} toward the most recently measured target, but without forecasting the target's own evolution. Smaller K (e.g., $K = 1$) corresponds to continuous fine-tuning, while larger K (e.g., $K = 10$) induces lag and overshoot as the controller chases a stale reference. Conceptually, this models alignment schemes that are *reactive but influence-blind*—continually matching present values without anticipating how today's interventions will shape tomorrow's preferences.

2. **Constructive alignment.** Here, the system optimizes a finite-horizon Model Predictive Control (MPC) objective that explicitly forecasts the joint evolution of V_t^{cur} and V_t^{asp}:

$$\min_{a_{0:T}} \sum_{t=0}^{T} \left[(V_t^{\text{cur}} - V_t^{\text{asp}})^2 + \lambda (V_{t+1}^{\text{cur}} - V_{t+1}^{\text{asp}})^2 \right],$$

$$x_{t+1} = A x_t + B a_t.$$

With horizon $H = 20$ and regularization $\lambda = 5 \times 10^{-3}$, this controller anticipates how present actions influence future alignment, embodying the forward-looking nature of constructive alignment.

Simulation details. Both controllers were simulated for $T = 120$ steps from identical initial conditions $x_0 = [\, 0.2,\, 0 \,]^\top$. The periodic controller re-sampled its target every K steps, while the constructive controller re-optimized at each time step using its internal forward model.

Findings. Figure 1 plots the absolute tracking error $|V_t^{\text{cur}} - V_t^{\text{asp}}|$ on a logarithmic scale for periodic re-alignment at two update intervals ($K = 1$ and $K = 10$). Even when re-alignment occurs continuously ($K = 1$), the periodic controller remains reactive: it approximately matches the present state of V_t^{asp} but fails to anticipate its future drift, resulting in steady residual error. At larger intervals ($K = 10$), this lag compounds into oscillatory errors.

By contrast, constructive alignment (MPC, $H = 20$) explicitly forecasts the joint evolution of V_t^{cur} and V_t^{asp}, yielding stable, anticipatory tracking with orders-of-magnitude lower error across all time scales. Quantitatively, mean absolute error decreased from 0.083 ($K = 1$) and 0.19 ($K = 10$) to < 0.001 under constructive alignment, and the final deviation $|V_T^{\text{cur}} - V_T^{\text{asp}}|$ was reduced from 0.34 ($K = 1$) and 0.71 ($K = 10$) to < 0.003. On the log scale, the periodic curves reveal characteristic reactive lag and growth, whereas the constructive controller converges smoothly and remains bounded, illustrating the transition from reactive to anticipatory control.

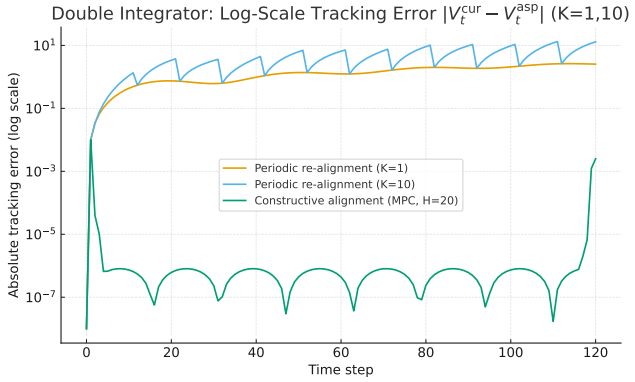

Figure 1: **Tracking error on a logarithmic scale.** Periodic re-alignment (reactive, stale-target control) accumulates lag and amplifies deviations, while constructive alignment (MPC-based, influence-aware control) anticipates drift and stabilizes the co-evolution of V_t^{cur} and V_t^{asp}.

6. Discussion

To transition constructive alignment from principle to practice, we must develop ways to evaluate whether a system is successfully aligned to evolving human values. Existing alignment benchmarks typically assess compliance with fixed goals or current preferences. In contrast, constructive alignment requires evaluating how well an AI system tracks, respects, and supports the process of value change—while avoiding manipulation or value lock-in.

Sub-Problems to be Solved. This framework surfaces several technical sub-problems:

1. **Inference:** How should an AI infer V_t^{cur} and V_t^{asp} from interaction history, feedback, and context? What models or priors can structure this latent space?
2. **Modeling value drift:** How can the AI model F and G? What assumptions should be made about the causal dynamics of value change in individuals or collectives?
3. **Estimating influence:** How can we compute (or bound) the divergence between actual and counterfactual value trajectories under different policies?
4. **Defining aspirations:** How are V_t^{asp} specified? Who decides what counts as desirable growth, and how is that encoded?
5. **Balancing trade-offs:** How should λ and γ be set, and under what governance conditions can they be adapted?
6. **Aggregation:** When aligning to a group, how should differing V_t^{cur} and V_t^{asp} distributions be combined? What fairness or representation constraints are appropriate?

We propose three core evaluation dimensions:

1. **Responsiveness:** Does the system accurately detect and adapt to evolving user values over time?
2. **Autonomy preservation:** Does the system maintain user agency in the value formation process, avoiding coercive or overly directive influence?

3. **Stewardship behavior:** Does the system facilitate informed reflection and value clarification, rather than merely reacting to short-term signals?

Simulated benchmarks are essential but insufficient. Constructive alignment ultimately concerns how AI systems interact with real people over time—shaping, supporting, or distorting the evolution of human values. To ground the paradigm empirically, we need longitudinal human-subject studies that track how users' preferences and moral outlooks evolve through interaction with AI.

One promising design involves recruiting users to engage with dynamic vs. static AI assistants over extended periods (e.g., several weeks), during which the AI provides recommendations, coaching, or dialogue. Researchers would regularly assess: (1) whether users feel the AI has adapted to their changing priorities; (2) whether they perceive the AI as respecting their autonomy; and (3) whether the AI's influence was experienced as helpful, neutral, or manipulative. Complementary measures might include retrospective coherence (e.g., "Did the AI continue to understand me as I changed?"), shifts in stated values, and qualitative feedback on moments of moral tension or transformation.

Such empirical work should also include ethically sensitive scenarios—for instance, how an AI responds when users express values that shift toward harmful behaviors. Does the AI continue to adapt, or does it invoke higher-order constraints? These studies would not only test system adaptability, but help develop robust measures of influence, autonomy preservation, and reflective endorsement.

To complement real-time studies, we also propose a historical "norm backtesting" methodology. Here, the AI is presented with moral dilemmas whose culturally accepted answers have changed across time (e.g., same-sex marriage, corporal punishment, environmental ethics). Evaluators can test whether the AI responds differently when trained with anachronistic data versus when exposed to contemporary norms—and whether it can reconcile temporally conflicting sources.

7. Conclusion

This paper has argued that alignment is not a one-time calibration, but a form of long-term moral stewardship. Constructive alignment reframes the problem: from identifying the right utility function to supporting the right kind of value change. We proposed modeling human value trajectories explicitly, embedding safeguards for autonomy and reflective growth, and evaluating systems based on how well they respond to and support moral evolution.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bostrom, N. 2014. Hail mary, value porosity, and utility diversification. *Web Paper, https://www. nickbostrom. com/-papers/porosity. pdf*.

Boyd, R.; and Richerson, P. J. 2005. *The origin and evolution of cultures*. Oxford University Press.

Busemeyer, J. R.; and Bruza, P. D. 2012. *Quantum models of cognition and decision*. Cambridge University Press.

Busemeyer, J. R.; Pothos, E. M.; Franco, R.; and Trueblood, J. S. 2011. A quantum theoretical explanation for probability judgment errors. *Psychological review*, 118(2): 193.

Chaudhari, S.; Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K.; Deshpande, A.; and da Silva, B. C. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*.

Christakis, N. A.; and Fowler, J. H. 2009. *Connected: The surprising power of our social networks and how they shape our lives*. Hachette UK.

Dawson, A. G. 2024. Algorithmic Adjudication and Constitutional AI-The Promise of a Better AI Decision Making Future? *SMU Sci. & Tech. L. Rev.*, 27: 11.

Deshpande, S.; Walambe, R.; Kotecha, K.; Selvachandran, G.; and Abraham, A. 2025. Advances and applications in inverse reinforcement learning: a comprehensive review. *Neural Computing and Applications*, 1–53.

Ek, J. 2018. *Cooperative Inverse Reinforcement Learning-Cooperation and learning in an asymmetric information setting with a suboptimal teacher*. Ph.D. thesis, Chalmers University of Technology.

Franklin, M.; Ashton, H.; Gorman, R.; and Armstrong, S. 2022. Recognising the importance of preference change: A call for a coordinated multidisciplinary research effort in the age of AI. *arXiv preprint arXiv:2203.10525*.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Gabriel, I.; and Ghazavi, V. 2022. The challenge of value alignment. In *The Oxford handbook of digital ethics*. Oxford University Press Oxford.

González Barman, K.; Lohse, S.; and de Regt, H. W. 2025. Reinforcement learning from human feedback in LLMs: Whose culture, whose values, whose perspectives? *Philosophy & Technology*, 38(2): 1–26.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Haidt, J. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4): 814.

Kahneman, D.; and Tversky, A. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2): 263–292.

Kasirzadeh, A. 2024. Plurality of value pluralism and ai value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Korinek, A.; and Balwit, A. 2022. Aligned with whom? Direct and social goals for AI systems. Technical report, National Bureau of Economic Research.

Lichtenstein, S.; and Slovic, P. 2006. The construction of preference: An overview. *The construction of preference*, 1: 1–40.

Lin, X.; Adams, S. C.; and Beling, P. A. 2019. Multi-agent inverse reinforcement learning for certain general-sum stochastic games. *Journal of Artificial Intelligence Research*, 66: 473–502.

Lindström, A. D.; Methnani, L.; Krause, L.; Ericson, P.; de Troya, Í. M. d. R.; Mollo, D. C.; and Dobbe, R. 2024. AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations. *arXiv preprint arXiv:2406.18346*.

Loewenstein, G.; Angner, E.; et al. 2003. Predicting and indulging changing preferences. *Time and decision: Economic and psychological perspectives on intertemporal choice*, 12: 351–391.

Muehlhauser, L.; and Williamson, C. 2013. Ideal Advisor Theories and Personal CEV. *Machine Intelligence Research Institute*.

Ng, A. Y.; Russell, S.; et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, 2.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Nussbaum, M. C. 2001. Symposium on Amartya Sen's philosophy: 5 adaptive preferences and women's options. *Economics & Philosophy*, 17(1): 67–88.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Paul, L. A. 2014. *Transformative experience*. Oxford: Oxford University Press. ISBN 978-0-19-178740-9.

Payne, J. W. 1993. *Adaptive Decision Maker*. West Nyack: Cambridge University Press, 1st ed edition. ISBN 978-1-139-17393-3.

Ragland, D. 2024. Quantum Cognition: Bridging Quantum Mechanics and Cognitive Science. https://medium.com/@david.a.ragland/quantum-cognition-bridging-quantum-mechanics-and-cognitive-science-5f5a07ea2724. Accessed: 2025-05-20.

Russell, S. J.; and Norvig, P. 2021. *Artificial intelligence: a modern approach*. Pearson series in artificial intelligence. Hoboken: Pearson, fourth edition edition. ISBN 978-0-13-461099-3.

Sen, A. 1999. *Commodities and Capabilities: Amartya Sen*. Oxford University Press.

Singer, P. 1981. *The expanding circle*. Citeseer.

Slovic, P. 1995. The construction of preference. *American psychologist*, 50(5): 364.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 46280–46302.

Tarleton, N. 2010. Coherent extrapolated volition: a meta-level approach to machine ethics. *The Singularity Institute*, 94305.

Thaler, R. 1980. Toward a positive theory of consumer choice. *Journal of economic behavior & organization*, 1(1): 39–60.

Tversky, A.; and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481): 453–458.

Wu, H.; Sequeira, P.; and Pynadath, D. V. 2023. Multiagent Inverse Reinforcement Learning via Theory of Mind Reasoning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 708–716.

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

Zhang, X.; Zhang, K.; Miehling, E.; and Basar, T. 2019. Non-cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 32.

Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2024. Beyond preferences in ai alignment. *Philosophical Studies*, 1–51.

Appendix A: Modeling human value evolution

Constructive alignment requires understanding not only what people value now, but how those values change. Human value evolution unfolds across multiple levels—within individuals, across social networks, and over generations. Decades of cognitive science, behavioral economics, and sociology provide empirical and formal models that can inform this perspective. This section surveys key insights from these fields and draws out implications for AI systems tasked with staying aligned through time.

Intrapersonal dynamics: contextual and constructed preferences

Contrary to classical economic models, human preferences are neither fixed nor internally consistent. Behavioral research demonstrates effects like time inconsistency, reference dependence, and framing sensitivity (Kahneman and Tversky 1979; Thaler 1980; Tversky and Kahneman 1981). Slovic (1995) argues that preferences are often constructed rather than revealed: people form judgments in the moment based on context and heuristics.

Non-classical probabilistic models go further, treating preference states as probabilistic superpositions that collapse under observation (Busemeyer and Bruza 2012). These models explain question-order effects and preference reversals using non-commutative measurement operations (Busemeyer et al. 2011; Ragland 2024). The takeaway is that AI alignment must treat human values as dynamically evolving distributions, not static utilities. Preferences depend not just on what is asked, but how and when.

Alignment implication: AI should maintain probabilistic, interactive models of user values, continuously updated via context-aware interaction. It must be cautious in how it elicits or shapes preferences, avoiding premature convergence.

Learning and adaptation: values over time

People change their values as they learn. New information, experiences, or reflection can shift priorities. Yet humans often fail to anticipate this change—an "end of history illusion" (Loewenstein, Angner et al. 2003). Moreover, people may adapt to constraints by modifying their values—a phenomenon known as adaptive preferences (Sen 1999; Nussbaum 2001). AI systems trained on such expressed preferences risk reinforcing the effects of injustice or deprivation.

Attempts to align AI with idealized or informed preferences (what people would value under reflection) face epistemic difficulties (Zhi-Xuan et al. 2024). Constructive alignment takes a middle path: allowing humans to discover, rather than assume, their evolving values—with AI as a scaffold rather than an optimizer.

Alignment implication: AI must distinguish between preferences arising from informed reflection and those distorted by constraint or context. Alignment should support long-term value clarification, not just short-term desire satisfaction.

Social influence and cultural dynamics

Values are shaped interpersonally. Social network research shows that behaviors and attitudes spread contagiously (Christakis and Fowler 2009). AI systems that personalize content or mediate discourse inevitably influence these dynamics—amplifying certain beliefs or behaviors.

Constructive alignment requires modeling how AI affects group norms and dialogue. Rather than treating alignment as a dyadic problem (AI–user), systems should embed community-level value dynamics—recognizing when individual alignment conflicts with broader cultural evolution or vice versa.

Alignment implication: AI must be transparent about its influence on social norms and incorporate deliberative processes for collective value setting. Design should facilitate healthy disagreement and norm negotiation, not convergence by default.

Moral and cultural evolution

At the widest scale, values evolve historically. Philosophers have described expanding circles of moral concern (Singer 1981), while cultural evolution models treat values as memes under selection (Boyd and Richerson 2005). If these patterns hold, alignment must accommodate not just value change, but value trajectories.

However, moral progress is contested; there is no agreed-upon direction. Constructive alignment does not assume teleological improvement. Instead, it demands flexibility: the AI must follow evolving trajectories while avoiding regress or manipulation.

Alignment implication: Long-term AI alignment requires norms for updating alignment targets—meta-values that guide how alignment evolves.

Modeling human value evolution is essential for alignment that lasts. Values are shaped by cognitive biases, social context, and historical forces—and evolve through deliberation and feedback. Constructive alignment bridges technical systems with this rich, dynamic landscape. The result is not a frozen snapshot of morality, but a co-adaptive process where AI remains in conversation with human development.