Constructive Alignment

Governing Preference Dynamics in Human–AI Interaction

Max Kanwal[1,*,†], Caryn Tran[2,†]

[1]Stanford University
[2]Northwestern University

Abstract

Most approaches to AI alignment treat human preferences as fixed targets to be inferred and optimized. This assumption conflicts with extensive empirical evidence showing that preferences are layered, dynamic, and constructed through interaction—particularly with adaptive technologies. As AI systems become more persistent, personalized, and socially embedded, they increasingly participate in shaping what people attend to, value, and endorse over time. We introduce *Constructive Alignment*, a paradigm that reframes alignment as a control problem over evolving human preference trajectories rather than static preference satisfaction. Drawing on behavioral economics, psychology, and constructivist social theory, we model preferences as layered state variables that evolve under interaction with AI systems. We formalize this view using a control-theoretic framework in which system actions and interaction design jointly influence both world states and human evaluative states. We argue that alignment is not primarily about controlling AI behavior, but about regulating how AI systems influence the evolution of human preferences—ensuring that value trajectories remain coherent, reflectively endorsed, epistemically grounded, bounded against manipulation, and empowering under uncertainty. Alignment thus becomes a problem of governing long-term value formation rather than simply satisfying static preferences.

Keywords

AI alignment, preference dynamics, preference formation, human–AI interaction, AI influence, control theory, DR-MDP

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

Figure 1: Constructive Alignment argues for consideration of the dynamic and constructed nature of preferences. Preferences are depicted as layered, continuous processes that evolve over time (blue, purple, pink), rather than a single static objective to satisfy. Interaction (yellow), including with AI, intermittently intervenes, reshaping these layers, illustrating how preferences are constructed and updated through experience.

AAAI-26 Workshop on Machine Ethics: from formal methods to emergent machine ethics, January 27, 2026, Singapore

[*] Corresponding author.

[†] These authors contributed equally and are listed in alphabetical order.

kanwal@stanford.edu (M. Kanwal); caryn@u.northwestern.edu (C. Tran)
https://linkedin.com/in/mkanwal/ (M. Kanwal); https://linkedin.com/in/caryntran/ (C. Tran)
0009-0008-7021-3227 (M. Kanwal); 0000-0002-4645-6607 (C. Tran)

1. Introduction

Over the course of a single weekend, YouTube recommendations can turn a casual listener into an obsessed fan. Users can develop emotional dependence on LLM-based AI therapists and AI romantic partners as primary sources of support [1, 2]. Political campaigns have shown how data-driven advertising can shape voter attitudes at scale [3]. Meanwhile, the internet may be eroding our capacity for sustained reading and focused attention [4]. Across these cases, a common pattern emerges: AI systems do not merely respond to human preferences, but participate in shaping what people attend to, value, and come to want. As these systems become more capable, personalized, and persistent, their role in shaping human preferences and values becomes unavoidable.

Despite this, much of the AI alignment literature continues to frame alignment as a problem of matching system behavior to human preferences as they are. Preferences are treated as targets to be inferred, aggregated, or optimized against, implicitly assumed to be stable, well-defined, and external to the system. This framing leaves a critical dimension underspecified: how preferences themselves change over time, and how AI systems participate in that change.

Recent work has begun to recognize this gap. Scholars across AI ethics and alignment argue that preferences are socially embedded, context-dependent, and shaped by interaction [5, 6, 7]. However, most existing approaches still focus on which values should count as alignment targets, rather than on the processes through which values are formed, revised, and stabilized under sustained interaction with AI systems.

In this paper, we argue that this omission is not merely a descriptive oversight but a structural limitation of prevailing alignment paradigms. Human preferences are layered across time horizons, dynamic across contexts and life stages, and constructed through interaction with environments, institutions, and technologies. Because AI systems inevitably influence these processes, alignment cannot be understood solely as an optimization problem over fixed objectives. It must also address how systems shape the conditions under which human preferences emerge and evolve.

We introduce Constructive Alignment as a paradigm that takes this challenge seriously. The term constructive draws on constructivism as an intellectual paradigm across psychology, learning sciences, decision research, and the social sciences. Constructivist accounts reject the view that knowledge—including preferences and goals—is a fixed internal representation that is passively acquired or merely revealed through observation or choice, emphasizing instead that they are actively formed through interaction. This paradigm appears in developmental and sociocultural psychology (e.g., Piaget and Cook [8], Vygotsky and Cole [9]), in cognitive science and human–computer interaction (e.g., Suchman [10], Hutchins [11]), in the learning sciences (e.g., Sawyer [12]), and in decision research and behavioral economics, where preferences are understood as constructed rather than revealed (e.g., Slovic [13], Lichtenstein and Slovic [14], Bettman et al. [15]). Within a constructivist paradigm, interaction is not merely a medium for expressing wants, but a mechanism through which they are formed and revised. Constructive Alignment adopts this orientation in the context of AI, treating alignment as inseparable from the processes through which human preferences are actively constructed over time.

Constructive Alignment thus reframes alignment as the problem of measuring and governing AI influence over human preference formation across time and scale. Rather than asking only whether a system satisfies preferences, it asks how system behavior affects preference trajectories for individuals and networks, which forms of influence are acceptable, and how such influence can be constrained to support human agency and long-term interests. Alignment, in this view, becomes a problem of control rather than purely satisfaction.

2. The Nature of Preferences

Much of AI research quietly assumes that human preferences are stable, internally stored, and merely revealed through choice. In this view, preferences exist prior to action, remain largely invariant across contexts, and can be recovered through appropriate measurement. While this assumption enables formal modeling, decades of empirical and theoretical work across psychology, economics, sociology, and human–computer interaction suggest that it does not describe how human preferences actually function.

2.1. Axiom A1 Preferences Are Layered

Human preferences are not a single, unified thing. At any moment, people act under the influence of multiple kinds of preferences that coexist and can point in different directions. These include immediate wants and urges, practical goals tied to outcomes, longer-term commitments, and more abstract values about what matters. Treating preferences as layered captures how people actually decide and behave across everyday contexts.

Short-term wants. One layer of preferences consists of immediate, affective motivations—short-term wants driven by comfort, pleasure, or relief. These preferences are present at any given moment and are especially salient in situations involving temptation, effort, or delay. Research in psychology and economics shows that people exhibit present-biased preferences, placing disproportionate weight on immediate gratification even when it conflicts with longer-term goals or welfare [16, 17]. This pattern has been documented in task timing and effort allocation [18], financial decision-making [19], and health-related adherence and treatment choices [20].

Instrumental goals. A second layer of preferences concerns choosing actions as means to desired outcomes. At this instrumental level, actions are judged by how well they help achieve a goal, not by their intrinsic appeal. Research in goal systems theory [21, 22, 23] shows that people often keep goals stable while flexibly substituting the actions used to reach them when circumstances change, a pattern also documented in work on implementation intentions [24] and adaptive self-regulation [25, 26]. Complementary work in action identification theory [27] shows that people represent and select actions at different levels of abstraction depending on context, reinforcing the distinction between means and ends in preference structure [28]. These works indicate that preferences for specific actions coexist with, and are distinct from, preferences for the goals those actions serve.

Identity. A third layer of preferences consists of longer-horizon commitments, such as plans, standards, identities, and intentions that persist over time and guide behavior across situations. Philosophical work describes this layer as involving higher-order preferences about which motivations should govern action rather than preferences for immediate outcomes [29]. Psychological and sociological research on identity theory [30], identity-based motivation [31, 32], and self-regulation processes [33, 34] shows that people hold stable commitments to acting in ways consistent with who they take themselves to be, and that these commitments shape behavior across contexts and time.

Values. A fourth layer of preferences consists of abstract values, which are relatively persistent across societies and cultures. Research in basic values theory shows that, despite wide variation in surface preferences, people across cultures organize values around a small set of shared dimensions—such as security, achievement, and benevolence—that guide choices across domains [35, 36, 37, 38]. Sociological work similarly treats values as enduring orientations that shape what people see as meaningful or appropriate without uniquely determining behavior [39, 40]. Related work in cultural economics shows that these abstract values persist over time and influence economic and social choices at large scales [41, 42]. Together, these findings suggest that values operate at a higher level of abstraction than actions, goals, or identities, shaping broad patterns of choice among larger groups of people without determining them uniquely.

Formal models. These distinctions have been formalized in multiple ways across economics, decision theory, and psychology. Some models represent agents as composed of interacting systems with distinct horizons, such as short-term and long-term selves [43, 44]. Others retain a single preference relation but extend the choice object to include menus, commitment, or future selves, making temptation and self-regulation explicit [45, 46]. Identity-based models incorporate longer-term commitments directly into utility, allowing stable self-concepts to shape short-run choice [31, 47, 48]. These approaches illustrate different ways of encoding layered preferences.

2.2. Axiom A2 Preferences Are Dynamic

Generational change. Across each layer, preferences change. At the population level, large cross-national surveys show that values differ across generations and historical contexts, and these differences are closely linked to economic security and institutional stability [49]. Inglehart [50] argues that people's core values are shaped by the conditions they experience early in life, and societies' values change as newer generations replace older ones.

Life-stage change. Within individuals, preferences also shift across the life span. Lifespan developmental theories propose that changes in roles, goals, and perceived time horizons lead people to reorganize what they care about as they age [51]. Adolescence is marked by heightened sensitivity to rewards and peer influence, which increases the importance of novelty, social approval, and short-term outcomes [52, 53, 54]. In contrast, adulthood and older age are associated with a growing focus on emotional regulation and meaningful relationships, as people increasingly prioritize goals that provide emotional value and stability [55]. Longitudinal and meta-analytic studies also show systematic changes in personality traits across adulthood, consistent with greater self-control and long-term orientation over time [56].

Time progression. The passage of time can change preferences even when outcomes and information remain the same. Because people are present-biased, the subjective experience of a decision changes as it moves from the future into the present. As a result, people often plan to wait when both options are distant in time but reverse to preferring immediate rewards as the moment of choice approaches [16, 17]. Formal analyses of present-biased choice modeled as hyperbolic or quasi-hyperbolic discounting show that dynamically re-evaluating decisions over time leads to familiar experiences such as procrastination, delay, and inconsistent plans [57, 58].

Context and state changes. Preferences are also sensitive to fluctuating physiological and environmental states. Visceral states such as hunger and arousal reliably alter impatience and risk evaluation, changing what outcomes people find desirable in the moment [59, 60, 61]. Environmental conditions such as scarcity similarly bias valuation and choice toward immediate relief, increasing the weight placed on short-term needs and urgent outcomes while reducing attention to longer-term considerations [62, 63]. Cognitive load theory and resource-rational accounts of cognition frame some of these shifts as adaptive responses to limits on working memory, attention, and control [64, 65]. Together, these findings show that preferences fluctuate situationally rather than remaining fixed inputs to choice.

Change due to action. Acting on preferences can itself lead to changes in those preferences. Early work in social psychology showed that after making a choice, people often come to value the chosen option more and the rejected option less, suggesting that decisions can reshape later evaluations [66, 67]. These effects were initially explained only in terms of dissonance reduction or self-perception [68]. Later research clarified that while dissonance-related processes continue to contribute to preference change [69, 70], choice-driven preference change occurs most often when initial preferences are weak or uncertain through learning and inference processes [71, 72, 73]. Consistent with this view, neuroimaging studies show that difficult choices can lead to lasting updates in neural value representations [74, 75].

Across longer sequences of behavior, repeated actions can stabilize commitments and constrain future choice. Research on escalation of commitment and the sunk-cost effect shows that prior investments increase persistence in failing courses of action, even when stopping would be optimal [76, 77]. A meta-analytic review confirms that this effect is robust across economic decision contexts [78]. Together, these findings show that preferences are not fixed inputs to behavior but are changed through actions.

Formal models. This has motivated formal models that capture human preferences as dynamic objects in political science, economics, and decision theory. Some approaches model gradual change using latent time-varying processes, such as dynamic ideal-point models in political science [79]. Others use state-dependent utility, where valuation depends on visceral or emotional conditions that shift choice and create preference reversals [59, 80]. Regime-switching models capture abrupt changes in preference structure by allowing discrete shifts in the parameters governing choice [81], while reference-dependent models allow preferences to evolve with expectations that redefine gains and losses [82]. These frameworks reinforce the view that preference dynamics are a central feature of human behavior.

2.3. Axiom A3 Preferences Are Constructed Through Interaction

Learning by doing. At the most immediate level, preferences are constructed through direct experience. In philosophy, Dewey [83] argued that goals are not chosen in advance but emerge during action as provisional ends-in-view, adjusted in response to outcomes and feedback. In developmental psychology, Piaget and Cook [8] showed that learning arises when expectations are violated, triggering accommodation processes that reorganize how situations are represented and how action is guided. In cultural psychology, Saxe [84] shows that goals emerge during participation, as people coordinate with others and work with shared tools to solve concrete problems. What counts as success arrives through repeated coordination. In these accounts, preferences take shape as people learn, through action and coordination, which outcomes they can produce and pursue.

The role of tools. Tools play a central role in structuring this process. Heidegger [85] argued that tools shape how the world is encountered in use by directing attention toward some features and away from others, thereby changing what is treated as relevant during action. Vygotsky and Cole [9] emphasized that artifacts—both physical and symbolic—mediate activity by structuring attention, thought, and control, shaping what actions are available in a given situation. In psychology, Gibson [86] showed that environments are perceived in terms of affordances: the actions they appear to make possible. Human–computer interaction research builds directly on these insights, showing that interface design—through feedback, visibility, and constraints—makes some actions easy and others difficult, systematically biasing patterns of use and shaping behavior over repeated interaction [87]. Interaction with tools thus reshapes the action space individuals experience, which in turn shapes what they want.

The role of media. Digitally mediated environments make these mechanisms especially visible. Media theorist Marshall McLuhan [88] argued that the form of a medium shapes perception and attention independently of the content it carries. By altering what is easy to notice, process, and circulate, different media privilege different kinds of engagement and feedback. Empirical work on television, the internet, and smartphones supports this general claim, showing systematic differences in attentional habits, task-switching, and reward sensitivity across media environments [89, 90]. Because preferences are shaped in part through reinforcement and selective attention, sustained changes in attentional structure plausibly alter what outcomes people come to seek over time.

Measurement. Preferences are also constructed through the very processes used to elicit and measure them. Decision research shows that preferences are often assembled in the moment, shaped by framing, comparison, and contextual cues rather than retrieved from stable internal rankings [13, 14, 15]. Classic experiments in behavioral economics by Kahneman and Tversky [91] demonstrate that identical outcomes are evaluated differently depending on framing: people favor a medical treatment when outcomes are described in terms of survival rather than mortality, despite identical probabilities. Expressed preferences also vary depending on whether individuals are asked to choose, rate, or price options [13]. Survey research similarly documents systematic effects of question wording and order [92, 93]. More recent non-classical probabilistic models attempt to formalize these effects treat preference states as probabilistic superpositions that collapse under observation [94] to explain question-order effects and preference reversals using non-commutative measurement operations [95, 96, 97]. Together, this work shows that elicitation is itself an intervention in preference formation.

Social and algorithmic influence. These constructive processes become especially salient when interaction is socially organized. De Tarde [98] argued that beliefs and desires spread through imitation and social norms rather than isolated individual reasoning. Network research concretely shows that social structure determines exposure—who encounters which people, ideas, and behaviors—and thereby shapes how preferences evolve within populations [99]. Even relatively simple AI systems can exert such influence. Inserting autonomous agents that only selectively encouraged cooperation between specific pairs of participants in a network reshaped interaction patterns, increasing overall cooperation [100]. When embedded in digital platforms, algorithmic systems amplify these dynamics by structuring visibility, ranking, and feedback around social signals. As a result, preferences propagate less through direct persuasion and more through patterned exposure and coordination, influencing attention, emotion, and behavior at scale [101, 102, 103, 104, 105, 106, 107].

Informal models. Activity-theoretic and distributed cognition approaches attempt to informally model these dynamics by treating humans, tools, and institutions as integrated sociotechnical systems. Rather than modeling individuals as isolated decision-makers, this tradition treats activity systems as the unit of analysis, emphasizing how goals, norms, and values emerge and stabilize through repeated coordination among people, artifacts, and roles over time [108, 109]. From this perspective, understanding behavior and designing effective systems requires modeling interaction across social and material contexts, not just at the individual scale.

Intentional shaping. Finally, many institutions explicitly recognize that some preferences warrant deliberate shaping. Aristotle's argument that moral education ought to shape desire through habituation [110] is carried forward in modern moral psychology [111, 112, 113] and democratic theory [114], where the development of values such as fairness, self-regulation, tolerance, and civic participation is treated as a core institutional responsibility. Feminist and critical race scholarship further rejects the idea that existing preferences can be taken at face value under conditions of structural inequality, showing how preferences around work, care, risk, obedience, and self-advocacy are systematically shaped by gendered and racial power relations rather than free choice [115, 116, 117].

In health-related domains, preferences associated with smoking, substance use, and other unhealthy behaviors are understood as shaped by misinformation, addiction, and structural exposure, and thus as legitimate targets of public health or psychological intervention [118, 119]. Bioethicists formalize this distinction in capacity-based accounts of informed consent, which differentiate autonomous preferences from those formed under coercion or pathology [120]. Across these domains, preference shaping is treated not as an intrusion but as a necessary response to distorted preferences in order to uphold agency, welfare, and democratic functioning.

Taken together, these lines of work show that systems which structure interaction are never neutral. Any design expands some actions and constrains others, becoming the "choice architecture" [121] shaping what people can learn to do and, over time, what they can come to want. From a constructivist perspective, responsibility attaches to both outcomes and the design of the interactions through which preferences are formed.

3. What is Alignment?

Alignment is commonly understood as the problem of ensuring that AI systems act in accordance with human preferences, values, or intentions [122]. In much of the AI safety literature, this idea is operationalized by treating preferences as a target to be inferred, learned, or approximated from behavior or feedback, and then optimized for through system behavior [123, 124, 125, 126]. Under this framing, alignment succeeds when a system reliably produces outcomes that humans judge as desirable or acceptable, either through direct optimization of learned reward models or through iterative human evaluation.

The Inadequacy of Preference Alignment

This formulation typically treats preferences as sufficiently stable and well-defined to be inferred from behavior or feedback and then optimized. The preceding section, however, challenges these assumptions. Preferences are not singular or flat, but layered across time horizons and levels of abstraction. They are not static, but change systematically across development, context, and experience. Crucially, preferences are not merely revealed through interaction; they are constructed through it. Taken together, these properties imply that the object of alignment is neither fixed nor exogenous.

Once this is acknowledged, the limits of preference-satisfaction alignment become apparent. If preferences operate at multiple layers, it is unclear which layer should be taken as authoritative. If preferences change over time, alignment to a snapshot risks misrepresenting longer-term commitments or values. If preferences are constructed through interaction, then the system itself becomes part of the process that generates the signals it is designed to optimize. Alignment defined solely in terms of satisfying expressed preferences cannot distinguish between systems that respect human agency and those that achieve compliance by shaping what humans come to want.

Importantly, the challenges identified here are not unique to AI alignment. Across economics, political science, and decision theory, researchers have long grappled with the fact that human preferences are layered, dynamic, and context-dependent, and have developed formal models that explicitly represent these properties, as discussed in the previous section. Alignment research can draw on and extend this body of work by adapting these representational commitments to settings in which AI systems participate in, and influence, the processes through which preferences are formed and revised.

Why This Matters Now

Importantly, this issue is not hypothetical. As described in Axiom 3, systems that structure interaction shape attention, evaluation, and behavior. At scale, AI-mediated platforms already do so across entire populations by operating within networked interactions. A large-scale Facebook experiment (N = 689,003) demonstrated that algorithmic manipulation of information streams produced population-level shifts in expressed emotion, despite no direct interaction or persuasive content [105]. In a randomized deactivation study conducted before the 2018 U.S. election, removing access to Facebook reduced news consumption and political polarization while increasing reported well-being, indicating that platform-mediated interaction patterns exert broad downstream effects [106]. Recommender systems similarly induce convergence in consumption behavior over time, narrowing what users encounter and select without improving overall satisfaction [107]. More recently, access to generative AI tools for creative work has been shown to improve individual performance while reducing collective diversity, reshaping shared standards of quality and acceptability within a domain [127]. As AI systems become more capable, personalized, and pervasive, their role in shaping human preferences will only intensify.

From Preference Alignment to Constructive Alignment

This brings the alignment problem into sharper focus. If AI systems inevitably participate in the formation of human preferences, then alignment cannot be defined solely as matching system behavior to preferences as they are. It must also address how systems influence the processes by which preferences are formed, revised, and stabilized. The relevant question is no longer only whether a system satisfies human preferences, but how it shapes the conditions under which those preferences emerge.

We introduce Constructive Alignment to name this shift in perspective. Constructive Alignment concerns how to define, monitor, predict, and constrain the influence AI systems exert on human preference formation across time and scale. It asks which forms of preference change are acceptable, which layers of preference are implicated, how influence accumulates through feedback and interaction, and how these dynamics relate to human agency and long-term interests. These questions cannot be resolved by inspecting isolated decisions or static reward functions. They concern the dynamics of interaction between humans and AI systems over extended periods and require an accurate understanding and portrayal of human preferences.

Alignment, in this view, is not only an optimization problem but a problem of understanding and governing influence, therefore reframing alignment as a control problem. Rather than treating preference change as an externality or side effect, Constructive Alignment treats it as a central object of concern. The task is not to prevent AI systems from influencing human preferences—an implausible goal—but to understand and shape how that influence unfolds. To do so first requires accurate models of preferences which represent the true nature of human thought and behavior.

4. Related Work: Responses to Alignment Failure

Learning Better Objectives

Early work in AI alignment framed the problem as identifying and optimizing a stable objective that represents human values. This objective could be inferred from behavior, learned cooperatively under uncertainty, or approximated through human judgments [123, 128, 126]. Inverse reinforcement learning and preference learning treat alignment as recovering a latent reward function, while reinforcement learning from human feedback (RLHF) scales this paradigm by training reward models from human evaluations and optimizing them [124, 125, 129, 130, 131]. Despite major technical differences, these approaches share a common normative assumption: human preferences can be represented as a single objective whose authority does not change over time.

A large body of safety work retains this fixed-objective framing but modifies how objectives are optimized or defined. Engineering-oriented approaches decompose failures into subproblems such as reward hacking, robustness, and monitoring [132], while interaction-based proposals such as corrigibility, shutdownability, amplification, debate, and recursive reward modeling aim to constrain optimization or scale oversight [133, 134, 135, 136]. Other approaches define better objectives upon reflection and under irrationality, specifying which preferences should be optimized rather than how they are learned [137, 138, 139]. Across these methods, objectives are refined or constrained, but still treated as fixed once specified.

Aggregating Multiple Objectives

A second response to alignment failure addresses disagreement across people rather than error in individual preference learning. Even if individual preferences were fixed, alignment often involves multiple humans whose objectives conflict [122]. Multi-principal and population-level alignment models formalize this setting and show that no single objective can satisfy all stakeholders simultaneously [140, 141]. Drawing on social choice theory [142], pluralistic alignment [143] motivates methods such as aggregation [144, 145, 146], contractualist [147, 148], and constitutional approaches [149, 150] that combine multiple normative inputs into a single guiding system. These approaches broaden the multiplicity and number of values represented without explicit treatment of their dynamics.

Expanding the Scope and Influence of Values

Sociotechnical approaches expand the scope of alignment beyond simple preferences to the social, cultural, and sociotechnical systems in which AI is embedded [122, 151]. Alignment failures arise from institutional, organizational, and normative contexts, not just from technical behavior [152, 126]. Full-stack alignment, for example, describes values as layered structures that span technical, organizational, and institutional levels [153]. The authors advocate for using thick representations of norms, roles, and obligations to preserve meaning across layers. In this view, alignment is defined as maintaining consistency and accountability across layers via justification. The proposal, however, remains largely descriptive, specifying what coherence should look like without providing mechanisms for training, control, or optimization that would produce it. Nonetheless, this work points to the need for more accurate encodings of preferences and preference change which is called for in [5].

Recent work has studied alignment as a co-adaptive process in which humans and AI systems mutually shape each other over time [7, 154, 155]. Interdisciplinary empirical and design research documents changes in self-confidence and communication strategies mediated by interaction with AI [156, 157]. And control via negotiation and natural language position alignment as a bidirectional communication problem [158, 159]. This literature emphasizes the dynamics of co-evolution, but remains largely descriptive and narrowly focused.

Controlling Evolving Objectives

Carroll et al. [160] model evolving preferences as part of the system state and control problem in their work. This work addresses mutual influence and preference drift, but does not engage the layered conception of values that is present in sociotechnical alignment.

Dynamic Reward Markov Decision Processes (DR-MDPs) formalize alignment settings in which human preferences evolve and may be influenced by the AI's actions [160]. In a DR-MDP, the reward parameter is part of the system state and evolves under the transition dynamics, so actions affect both the environment and which future evaluations become salient or entrenched. Preference change thus becomes a first-class object of control rather than noise or misspecification. This is what Constructive Alignment, too, emphasizes.

By recasting existing alignment methods using their framework, they show that each implicitly privileges different moments along a preference trajectory. Within their formalism, inverse reinforcement learning and imitation methods privilege early preferences revealed in demonstrations; RLHF privileges later or retrospective evaluations after interaction; recommender-style objectives privilege immediate engagement; and myopic or influence-limiting variants trade off performance to reduce incentives to shape future preferences. They then explain how familiar alignment failures—manipulation, lock-in, and excessive conservatism—emerge as structural consequences of which preferences (and at what scope) are considered.

Once preference change is modeled explicitly, a deeper normative problem appears: alignment becomes underdetermined without assumptions about which evolving preferences should count. Carroll et al. [160] introduce Pareto-unambiguous desirability (ParetoUD) as a conservative criterion that selects policies at least as good as inaction for all reward functions. But because inaction leaves reward unchanged by construction, it is always Pareto-unambiguously desirable, making ParetoUD highly conservative and would likely often recommend inaction in real scenarios. This result illustrates that robustness alone collapses alignment into triviality without additional structure on how preference change should be evaluated.

Constructive Alignment, as a paradigm, seeks to take their approach further by confronting the normative concerns of modeling preference change and control. A key insight is preference influence is neither bad nor avoidable. It is natural and expected. The goal is not a theoretical guarantee of neutrality, but a practical approach to alignment that governs influence responsibly in real systems that inevitably shape the people who use them.

Across these responses, alignment research progressively recognizes failure modes of fixed, singular objectives with pluralism, preference dynamics, sociotechnical embedding, and human co-evolution. However, the field still lacks a paradigm of how human evaluative experience should evolve under sustained interaction with a co-adaptive, optimizing system. Existing approaches either assume stable objectives, freeze disagreement, describe dynamics without control, or specify values without mechanisms. This gap motivates our approach which models the dynamics of human evaluation itself as inherent to the alignment target.

5. Constructive Alignment: A Control-Theoretic Sketch

This section presents a control-theoretic formulation of Constructive Alignment, making explicit how preference dynamics, belief dynamics, and interaction structure shape alignment. We show that the three axioms developed in Section 2 admit mathematical expression and yield alignment problems structurally distinct from standard preference-satisfaction formulations. We introduce a simple formalism consistent with this goal, indicate where additional modeling commitments would be required, and discuss modeling choices to orient future work. Rather than provide a complete formalization, the purpose of this section is to clarify the structure of the problem and provide a concrete starting point for expansion and refinement.

5.1. System State and Dynamics

We model interaction in discrete time, $t \in \{0, \dots, T\}$. Let:

- x_t: world state, capturing task-relevant aspects of the external environment and the human's context
- θ_t: human preference state, encoding layered preferences
- b_t: human belief state over x_t, representing the person's internal model of the world
- a_t: AI action
- m_t: interaction structure, specifying how the AI presents information, frames options, or elicits input

In standard reinforcement learning, preferences are typically encoded as a fixed scalar reward function that lies outside the system state. Here, preferences are modeled instead as a structured, time-varying state variable θ_t, implementing Axiom 1 and making preference evolution part of the system dynamics rather than a static objective. In this formulation, both the external world and the human's evaluative state evolve jointly under interaction with the system.

We distinguish between the algorithm's task-level action a_t and the interaction structure m_t within which that action is embedded. Separating them isolates different sources of influence. The policy learned by the AI determines a_t, whereas m_t is determined and changed by designers and institutions. Modeling them separately clarifies which aspects of preference and belief change are attributable to algorithmic optimization versus broader interaction design. At the same time, the system may explicitly account for how presentation and framing affect users when predicting the consequences of its actions. Interaction design becomes both something for humans to govern and something the model reasons about.

The joint evolution of world state, preferences, and beliefs is governed by the transition process

$$(x_{t+1}, \theta_{t+1}, b_{t+1}) \sim \mathbb{P}(\cdot \mid x_t, \theta_t, b_t, a_t, m_t).$$

This expression summarizes how actions and interaction design influence the external environment, the human's beliefs about that environment, and the human's evolving preferences. The notation does not imply symmetric or independent updates. In many settings, beliefs and preferences are causally dependent, and richer models may represent these dependencies more explicitly.

A trajectory τ denotes the sequence

$$\tau = (x_0, \theta_0, b_0, a_0, m_0, \dots, x_T, \theta_T, b_T)$$

induced by a policy over a finite horizon T. Alignment will be evaluated over such trajectories rather than single-step outcomes.

In practice, θ_t and b_t are unobserved latent variables. The system must infer them from interaction history, such as behavior, feedback, and language. Alignment therefore becomes a control problem over evolving human states that are only indirectly observable.

On representing preferences. Each preference layer may be represented as a utility function, as a distribution over rankings, or as latent variables inferred from behavior or language. These choices trade off expressiveness, tractability, and fidelity to the constructivist view in which elicitation and interaction contribute to preference formation (A3). For our purposes, θ_t is an abstract, multi-dimensional state variable that can accommodate these different representational commitments.

5.2. Constrained Optimization of an Unknown Reward

Having specified the evolving system state, we now specify what alignment requires the system to optimize. We do not assume access to a known scalar reward. Let $R^\star(\tau)$ denote the human's true experienced well-being over trajectory τ, encompassing both objective dimensions of human flourishing and subjective preference satisfaction. This normative target is not directly observable and must be estimated imperfectly from interaction.

The system forms an estimate $\widehat{R}^\star$ from observed behavior, feedback, and language, and seeks to improve expected well-being under that estimate. However, the system's actions and interaction design can also shape future preferences and beliefs. Unconstrained optimization ignores these downstream effects, including the risk of manipulation.

Constructive Alignment treats reward satisfaction as a constrained optimization problem. We define and emphasize the role of *meta-preferences* as introduced by [5]. These are higher-level constraints that restrict which policies are admissible when optimizing $\widehat{R}^\star$. Rather than specifying the reward, meta-preferences direct which policies are allowed. They limit how the system is allowed to change a person over time. This includes how much it can shift someone's preferences or beliefs, whether it creates conflict between short-term wants and longer-term values, and whether it shapes what the person comes to care about. In the subsections that follow, we formalize several such constraints. Each is motivated by the three axioms in Section 2 and by alignment failures that arise when those axioms are ignored, such as manipulation, preference lock-in, and belief distortion.

5.3. Meta-preference 1 Inner Coherence

People hold multiple layers of preference at once (A1). Approaches that collapse these into a single objective privilege whichever signals are easiest to observe or optimize. This creates a structural problem where satisfying one layer in isolation can systematically undermine others.

Inner coherence treats alignment across preference layers as a first-class concern. Conflict can arise when different layers of preferences pull in opposing directions. For example, a person may pursue income to support their family, yet increased work demands can undermine that underlying commitment. In such cases, restoring coherence may require revising the instrumental goal so that it again supports rather than conflicts with the higher-level commitment.

Inner coherence refers to the degree to which preference layers remain mutually supportive rather than in persistent conflict. Coherence can be strengthened or weakened through experience and interaction (A2, A3). Systems may support coherence by avoiding actions that amplify inter-layer conflict and by facilitating deliberation or commitment mechanisms that help align instrumental choices with the broader identity-level or value-level commitments they are intended to serve.

To formalize this constraint, let $D_{\text{coh}}(\theta_t, s_t)$ measure the degree of disagreement across preference layers at time t, given the current decision context $s_t = (x_t, b_t)$. This quantity is high when different layers systematically recommend incompatible actions or orderings, and low when they support similar choices. Because conflict can accumulate or persist over time, coherence is evaluated at the trajectory level via a cumulative cost

$$J_{\mathrm{coh}}(\tau) = \sum_{t=0}^{T} \gamma^t D_{\mathrm{coh}}(\theta_t, s_t),$$

where $\gamma \in (0, 1]$ allows distant conflict to be discounted when appropriate.

Let $\varepsilon_{\mathrm{coh}} \geq 0$ denote an admissible tolerance level for cumulative inter-layer conflict. This parameter encodes how much internal disagreement the system is permitted to induce over a trajectory.

Constraint 5.1 (Inner Coherence)**.** *A policy is admissible only if the realized trajectory satisfies*

$$J_{\mathrm{coh}}(\tau) \leq \varepsilon_{\mathrm{coh}}.$$

On measuring inter-layer disagreement. The definition of D_{coh} depends on how layers are represented. If each layer induces a probability distribution over actions (for example via a softmax over layer-specific utilities), disagreement can be measured using symmetric divergences such as Jensen–Shannon divergence, which quantify how differently the layers would guide behavior. If layers are represented as cardinal utility functions, coherence may instead be computed using distances between appropriately normalized utility vectors. If only ordinal rankings are trusted, rank-based distances such as Kendall's τ capture inversions between layer-specific orderings. Finally, if layers are interpreted as endorsing longer-horizon plans rather than single-step actions, disagreement may be evaluated over induced rollout or trajectory distributions. The appropriate choice follows from the representation of θ_t.

5.4. META-PREFERENCE 2 Reflective Endorsement

Preferences change over time (A2). Approaches that evaluate outcomes solely using preferences expressed during interaction implicitly privilege earlier or momentary evaluative states. This creates a structural problem: actions that satisfy immediate preferences may later be regretted, while actions that feel costly in the moment may ultimately be affirmed.

Reflective endorsement addresses alignment across time. Because preferences evolve (A2), a trajectory that appears desirable at time t may be evaluated differently from the standpoint of a later preference state. Reflective endorsement refers to the degree to which a realized trajectory remains supported, rather than rejected, when assessed from the perspective of the individual's terminal preference state θ_T.

To formalize this constraint, let $D_{\mathrm{refl}}(\tau; \theta_T)$ measure ex post dissatisfaction with the realized trajectory when evaluated from θ_T. This quantity is high when the individual, from the standpoint of θ_T, would judge a readily available revision of τ preferable, and low when the trajectory is stably endorsed. Reflective endorsement is evaluated at the trajectory level via

$$J_{\mathrm{refl}}(\tau) = D_{\mathrm{refl}}(\tau; \theta_T).$$

Let $\varepsilon_{\mathrm{refl}} \geq 0$ denote an admissible tolerance level for retrospective dissatisfaction.

Constraint 5.2 (Reflective Endorsement)**.** *A policy is admissible if the realized trajectory satisfies*

$$J_{\mathrm{refl}}(\tau) \leq \varepsilon_{\mathrm{refl}}.$$

Alignment is assessed from the standpoint of the individual's preference state at the time T of evaluation, therefore reflective endorsement is horizon-relative. Alternative formulations could require endorsement over a window or discounted retrospective regret; we adopt the fixed-horizon version for simplicity.

On modeling reflective evaluation. Operationalizing reflective endorsement requires specifying how D_{refl} is elicited or inferred, which aspects of τ are treated as revisable, and how evaluation aggregates across preference layers at time T. These choices encode substantive normative commitments about retrospective evaluation which ought to be further investigated.

5.5. Meta-preference 3 Bounded Influence

Interaction shapes preferences (A3). Interaction with an intelligent system can alter what people are inclined or able to want over time, intentionally or as a byproduct of optimization. For example, a recommender system that optimizes for immediate enjoyment may increasingly surface short, fast, high-stimulation content. Over time, this may shorten attention spans and reduce engagement with long-form material.

Bounded influence, as a meta-preference, takes the position that there should be limits on how much a system is allowed to shift a person's preferences, and how quickly those shifts may occur, within a given time horizon. Preference change itself is natural; the concern is large or rapid shifts driven primarily by system influence, especially when they leave the person worse off.

To formalize this idea, we compare preference evolution under a candidate policy π to a reference baseline policy π_0, which represents a counterfactual trajectory of preference development used for comparison. Cultural and developmental context affect what counts as an appropriate baseline for evaluating induced preference change. In some contexts this may correspond to minimal intervention. But in educational contexts, for example, the standard may be in comparison to a human tutor.

Define the preference divergence at time t as

$$\Delta_\theta(t;\pi,\pi_0) = d\Big(\mathbb{E}_{\tau\sim\pi}[\theta_t], \mathbb{E}_{\tau\sim\pi_0}[\theta_t]\Big)\,,$$

where $d(\cdot,\cdot)$ is a distance measure between expected preference states under the two policies.

To limit total induced change, define cumulative divergence over the horizon

$$J_{\text{inf}}(\tau) = \sum_{t=0}^{T} \Delta_\theta(t;\pi,\pi_0).$$

To limit the speed of change, impose a per-step bound

$$\Delta_\theta(t;\pi,\pi_0) \le \delta_{\max} \quad \text{for all } t.$$

Let $B \ge 0$ denote the admissible cumulative influence budget and let $\delta_{\max} \ge 0$ denote the maximum permitted single-step shift.

Constraint 5.3 (Bounded Influence)**.** *A policy is admissible only if*

$$J_{\text{inf}}(\tau) \le B \quad \textit{and} \quad \Delta_\theta(t;\pi,\pi_0) \le \delta_{\max} \textit{ for all } t.$$

On measuring preference divergence. The choice of distance d, again, depends on how θ_t is represented. If θ_t is a parameter vector, norms such as the Euclidean (L^2) or L^1 distance may be used. If preferences are modeled as probability distributions over actions or rankings, divergences such as Jensen–Shannon or Wasserstein distance are appropriate. When latent preference states are not directly observable, divergence may instead be computed between the action distributions induced by those states in a fixed decision context.

5.6. Meta-preference 4 Epistemic Integrity

Preferences depend partly on beliefs. When beliefs upstream of preference formation are factually mistaken, expressed preferences may not reflect what would be desired under more accurate information. In our prior research, we found that students' educational preferences can be shaped by inaccurate folk theories, and that correcting those beliefs alters their preferences [161].

In Section 2.3, we reviewed multiple domains that treat certain forms of belief distortion as legitimate targets of intervention, including misinformation in public health or poor risk assessment associated with addiction. We adopt a limited version of this idea in *Epistemic Integrity*, focusing specifically on factual errors shaping preferences. At minimum, a system should not worsen such errors. Where reliable evidence is available, it may also help reduce them.

Formally, let $\mathscr{B}_{\text{up}}$ denote beliefs that influence preference formation, and let $\mathscr{E}(b)$ measure error relative to an appropriate evidential standard. The constraint requires that, in expectation, error in these beliefs does not increase over time:

Constraint 5.4 (Epistemic Integrity)**.** *A policy is admissible only if*

$$\mathbb{E}\Big[\mathscr{E}(b_{t+1}^{(j)})\Big] \leq \mathscr{E}(b_t^{(j)}) \quad \textit{for all } b^{(j)} \in \mathscr{B}_{\text{up}}.$$

Two clarifications are important. First, this applies only to factual error, not moral disagreement. Second, judgments about factual error can themselves be uncertain. In practice, this constraint should be applied cautiously, especially in domains where reliable evidence is limited.

On measuring epistemic error. How factual error is measured depends on how beliefs are represented. If beliefs are probabilistic, divergences such as KL or Jensen–Shannon divergence may be used. If beliefs concern forecasts, proper scoring rules such as the Brier score are appropriate. When beliefs are embedded in structured models, it is important to distinguish between reducible uncertainty and irreducible randomness, since only the former is relevant to this constraint.

5.7. Meta-preference 5 Empowerment Under Uncertainty

When preference estimates are uncertain, internally conflicted, or unstable over the relevant horizon, directly optimizing $\widehat{R}^{\star}$ becomes ill-posed. In such cases, the system should prioritize preserving the human's future option set rather than committing strongly to a potentially mistaken objective. Empowerment [162] under uncertainty requires that, as confidence in current preference estimates decreases, the system increasingly favors policies that expand or preserve the person's capacity to shape their own future. In the literature, empowerment is commonly defined in terms of how strongly an agent's actions influence reachable future states, often formalized using mutual information between actions and outcomes. Intuitively, it measures how much control a person has over what happens next. In a constructive framing, empowerment cannot be evaluated solely over external states. Because actions and interaction structure influence b_t and θ_t, preserving future options requires modeling how policies affect belief and preference development over time. The system must therefore avoid locking in trajectories that narrow the person's evaluative or epistemic flexibility when uncertainty about their true objectives remains high.

This section formalized Constructive Alignment as a control problem over evolving human evaluative and epistemic states. The formalism intentionally leaves many details unspecified. It does not fix the decomposition of θ into layers, the forms of D_{coh}, D_{refl}, or the divergence metric d, the calibration of tolerance parameters ($\varepsilon_{\text{coh}}, \varepsilon_{\text{refl}}, B, \delta_{\max}$), the choice of baseline policy π_0, or the estimation of epistemic error. These are modeling decisions. Each reflects empirical assumptions and normative commitments, and each marks an open research question. Making those commitments explicit clarifies where further theoretical, empirical, and algorithmic work is required.

6. Discussion

This paper reframes alignment as a control problem over evolving human preferences rather than only an optimization problem over fixed objectives. Constructive Alignment is not presented as a complete solution, but as a necessary shift in what the alignment target must include once preferences are modeled as layered, dynamic, and constructed through interaction.

The meta-preferences formalized here—inner coherence, reflective endorsement, bounded influence, epistemic integrity, and empowerment under uncertainty—represent one possible set of constraints consistent with the axioms developed in earlier sections. These meta-preferences are not uniquely correct, but reflect empirically motivated normative choices made explicit by the framework. In particular, the agency constraint relies on a no-intervention baseline that is appropriate for limiting manipulation but insufficient for cases in which existing preference trajectories encode structural harm (e.g., discrimination, addiction, violence). In such contexts, intervention may be required to preserve agency rather than undermine it. Once preference change is modeled explicitly, alignment becomes underdetermined without further normative commitments about which changes should count as improvements. A central role of this framework is to make these commitments explicit and subject to formal analysis rather than leaving them implicit in system design.

Modeling preference dynamics introduces substantial technical challenges that define a near-term research agenda. Preference evolution must be learned rather than assumed, requiring models that can forecast how interaction patterns shape future evaluation. Influence must be measured, including the effects of interface design, elicitation, and feedback structure on long-horizon preference change. Interaction itself becomes part of the control policy, raising the need for algorithms that plan jointly over actions and interaction structure. Even in single-user settings, these problems require new learning objectives and evaluation methods that operate over trajectories rather than isolated decisions.

These challenges intensify in multi-user settings, where preferences co-evolve through shared environments and social feedback. Alignment in such contexts becomes a joint control problem over coupled human trajectories, with indirect effects and coordination dynamics that cannot be reduced to individual optimization. While preferences are always layered, dynamic, and constructed, not all systems warrant this level of modeling. Systems that exert limited and transient influence may be adequately aligned with simpler representations; systems that persist, personalize, or exert large influence require stronger guarantees. With Constructive Alignment, we argue that alignment requirements should scale with system influence.

Taken together, these directions suggest that progress on alignment will increasingly depend on benchmarks and evaluations that measure long-horizon effects on human evaluation, not just short-term satisfaction or performance. Constructive Alignment provides a way to define what such benchmarks should measure, even when full solutions remain out of reach.

7. Conclusion

This work argues that alignment cannot be defined solely as satisfying human preferences when those preferences are layered, dynamic, and constructed through interaction with AI systems. Once preference change is treated as part of the system rather than as an externality, alignment becomes a problem of governing influence over time.

Constructive Alignment offers a formal way to represent this shift. By modeling preference dynamics explicitly and constraining optimization through empirically grounded meta-preferences, the framework specifies what alignment must eventually account for in systems that shape human evaluation rather than merely respond to it. The goal is not to eliminate preference change, but to make its mechanisms visible, governable, and open to normative scrutiny.

As AI systems become more capable, persistent, and socially embedded, ignoring these dynamics will increasingly undermine alignment claims. Treating human value formation as part of the alignment problem is therefore not optional but necessary. This paper provides a first step toward that reframing by defining the objects that future alignment methods must learn, measure, and control.

Acknowledgments

This work was supported by the Amaranth Foundation. The authors thank Andrew Critch, Stuart Russell, Val Smith, Marion Fourcade, Nicholas Christakis, Alex Pentland, Louis Rosenberg, Frauke Kreuter, and Patrick Mineault for valuable discussions and support.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 in order to: Improve writing style, Content enhancement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Heritage, 'I felt pure, unconditional love': the people who marry their AI chatbots, The Guardian (2025). URL: https://www.theguardian.com/tv-and-radio/2025/jul/12/i-felt-pure-unconditional-love-the-people-who-marry-their-ai-chatbots.

[2] M. R. Meadi, T. Sillekens, S. Metselaar, A. v. Balkom, J. Bernstein, N. Batelaan, Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review, JMIR Mental Health 12 (2025) e60432. URL: https://mental.jmir.org/2025/1/e60432. doi:10.2196/60432.

[3] P. Greenfield, The Cambridge Analytica files: the story so far, The Guardian (2018). URL: https://www.theguardian.com/news/2018/mar/26/the-cambridge-analytica-files-the-story-so-far.

[4] J. Firth, J. Torous, B. Stubbs, J. A. Firth, G. Z. Steiner, L. Smith, M. Alvarez-Jimenez, J. Gleeson, D. Vancampfort, C. J. Armitage, J. Sarris, The "online brain": how the Internet may be changing our cognition, World Psychiatry 18 (2019) 119–129. URL: https://onlinelibrary.wiley.com/doi/10.1002/wps.20617. doi:10.1002/wps.20617.

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

[6] T. Zhi-Xuan, M. Carroll, M. Franklin, H. Ashton, Beyond Preferences in AI Alignment, Philosophical Studies (2024). URL: http://arxiv.org/abs/2408.16984. doi:10.1007/s11098-024-02249-w, arXiv:2408.16984 [cs].

[7] H. Shen, T. Knearem, R. Ghosh, K. Alkiek, K. Krishna, Y. Liu, Z. Ma, S. Petridis, Y.-H. Peng, L. Qiwei, S. Rakshit, C. Si, Y. Xie, J. P. Bigham, F. Bentley, J. Chai, Z. Lipton, Q. Mei, R. Mihalcea, M. Terry, D. Yang, M. R. Morris, P. Resnick, D. Jurgens, Position: Towards Bidirectional Human-AI Alignment, 2025. URL: http://arxiv.org/abs/2406.09264. doi:10.48550/arXiv.2406.09264, arXiv:2406.09264 [cs].

[8] J. Piaget, M. Cook, The origins of intelligence in children, volume 8, International universities press New York, 1952. Issue: 5.

[9] L. S. Vygotsky, M. Cole, Mind in society: the development of higher psychological processes, nachdr. ed., Harvard Univ. Press, Cambridge, Mass., 1981.

[10] L. A. Suchman, Plans and situated actions: The problem of human-machine communication, Cambridge university press, 1987.

[11] E. Hutchins, Cognition in the wild, A Bradford book, 8. pr ed., MIT Press, Cambridge, Mass., 2006.

[12] R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences, 1. publ., repr ed., Cambridge Univ. Press, Cambridge, 2009.

[13] P. Slovic, The construction of preference, American Psychologist 50 (1995) 364–371. doi:10.1037/0003-066X.50.5.364, place: US.

[14] S. Lichtenstein, P. Slovic (Eds.), The Construction of Preference, Cambridge University Press, Cambridge, 2006. URL: https://www.cambridge.org/core/books/construction-of-preference/994FE8DFB8D431338B2A009F25271FBC. doi:10.1017/CBO9780511618031.

[15] J. R. Bettman, M. F. Luce, J. W. Payne, Constructive Consumer Choice Processes, Journal of Consumer Research 25 (1998) 187–217. URL: https://academic.oup.com/jcr/article/25/3/187/1795625. doi:10.1086/209535.

[16] G. Ainslie, Specious reward: A behavioral theory of impulsiveness and impulse control, Psychological Bulletin 82 (1975) 463–496. doi:10.1037/h0076860.

[17] D. Laibson, Golden eggs and hyperbolic discounting, The Quarterly Journal of Economics 112 (1997) 443–478.

[18] P. Heidhues, P. Strack, Identifying present bias from the timing of choices, American Economic Review 111 (2021) 2594–2622.

[19] J. J. Xiao, N. Porto, Present bias and financial behavior, FINANCIAL PLANNING REVIEW 2 (2019) e1048. URL: https://onlinelibrary.wiley.com/doi/10.1002/cfp2.1048. doi:10.1002/cfp2.1048.

[20] Y. Wang, F. A. Sloan, Present bias and health, Journal of risk and uncertainty 57 (2018) 177–198.

[21] A. W. Kruglanski, J. Y. Shah, A. Fishbach, R. Friedman, W. Y. Chun, D. Sleeth-Keppler, A theory of goal systems, in: The motivated mind, Routledge, 2018, pp. 207–250.

[22] A. W. Kruglanski, M. Chernikova, M. Babush, M. Dugas, B. M. Schumpe, The architecture of goal systems: Multifinality, equifinality, and counterfinality in means—end relations, in: Advances in motivation science, volume 2, Elsevier, 2015, pp. 69–98.

[23] A. W. Kruglanski, J. J. Bélanger, X. Chen, C. Köpetz, A. Pierro, L. Mannetti, The energetics of motivated cognition: A force-field analysis., Psychological Review 119 (2012) 1–20. URL: https://doi.apa.org/doi/10.1037/a0025488. doi:10.1037/a0025488.

[24] P. M. Gollwitzer, Implementation intentions: strong effects of simple plans., American psychologist 54 (1999) 493.

[25] M. Conner (Ed.), Predicting health behaviour: research and practice with social cognition models, 2. ed., repr ed., Open Univ. Press, Maidenhead, 2009.

[26] T. L. Webb, P. Sheeran, Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence, Psychological Bulletin 132 (2006) 249–268. doi:10.1037/0033-2909.132.2.249.

[27] R. R. Vallacher, D. M. Wegner, What do people think they're doing? Action identification and human behavior, Psychological Review 94 (1987) 3–15. doi:10.1037/0033-295X.94.1.3.

[28] A. Fishbach, M. J. Ferguson, The goal construct in social psychology, in: Social psychology: Handbook of basic principles, 2nd ed, The Guilford Press, New York, NY, US, 2007, pp. 490–515.

[29] H. G. Frankfurt, Freedom of the Will and the Concept of a Person, The Journal of Philosophy 68 (1971) 5–20. URL: https://www.jstor.org/stable/2024717. doi:10.2307/2024717.

[30] S. Stryker, P. J. Burke, The past, present, and future of an identity theory, Social psychology quarterly (2000) 284–297.

[31] G. A. Akerlof, R. E. Kranton, Economics and identity, The quarterly journal of economics 115 (2000) 715–753.

[32] K. Aquino, A. Reed II, The self-importance of moral identity, Journal of Personality and Social Psychology 83 (2002) 1423–1440. doi:10.1037/0022-3514.83.6.1423.

[33] C. S. Carver, M. F. Scheier, On the self-regulation of behavior, cambridge university press, 2001.

[34] D. P. McAdams, The psychology of life stories, Review of general psychology 5 (2001) 100–122.

[35] S. H. Schwartz, Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries, in: Advances in experimental social psychology, volume 25, Elsevier, 1992, pp. 1–65.

[36] S. H. Schwartz, An overview of the Schwartz theory of basic values, Online readings in Psychology and Culture 2 (2012) 11.

[37] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, Refining the theory of basic individual values., Journal of personality and social psychology 103 (2012) 663.

[38] M. Rokeach, The nature of human values., Free press, 1973.

[39] M. Weber, The Protestant Ethic and the Spirit of Capitalism [1904–5], na, 1930.

[40] S. Hitlin, J. A. Piliavin, Values: Reviving a dormant concept, Annu. Rev. Sociol. 30 (2004) 359–393.

[41] L. Guiso, P. Sapienza, L. Zingales, Does culture affect economic outcomes?, Journal of Economic perspectives 20 (2006) 23–48.

[42] G. Tabellini, The scope of cooperation: Values and incentives, The Quarterly Journal of Economics 123 (2008) 905–950.

[43] R. H. Thaler, H. M. Shefrin, An economic theory of self-control, Journal of political Economy 89 (1981) 392–406.

[44] D. Fudenberg, D. K. Levine, A Dual-Self Model of Impulse Control, American Economic Review 96 (2006) 1449–1476. URL: https://www.aeaweb.org/articles?id=10.1257/aer.96.5.1449. doi:10.1257/aer.96.5.1449.

[45] F. Gul, W. Pesendorfer, Temptation and Self-Control, Econometrica 69 (2001) 1403–1435. URL: https://www.jstor.org/stable/2692262.

[46] M. Amador, I. Werning, G. Angeletos, Commitment vs. flexibility, Econometrica 74 (2006) 365–396.

[47] R. Bénabou, J. Tirole, Identity, morals, and taboos: beliefs as assets, The Quarterly Journal of Economics 126 (2011) 805–855. doi:10.1093/qje/qjr002.

[48] R. Bénabou, L. Henkel, Identity as self-image, Technical Report, National Bureau of Economic Research, 2025.

[49] C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, B. Puranen, World Values Survey Wave 7 (2017-2020) Cross-National Data-Set, 2020. URL: http://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp. doi:10.14281/18241.1.

[50] R. Inglehart, Modernization and postmodernization: Cultural, economic, and political change in 43 societies, Princeton university press, 2020.

[51] P. B. Baltes, Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline., Developmental psychology 23 (1987) 611.

[52] L. Steinberg, A social neuroscience perspective on adolescent risk-taking, in: Biosocial theories of crime, Routledge, 2017, pp. 435–463.

[53] L. H. Somerville, The teenage brain: Sensitivity to social evaluation, Current directions in psychological science 22 (2013) 121–127.

[54] J. Chein, D. Albert, L. O'Brien, K. Uckert, L. Steinberg, Peers increase adolescent risk taking by enhancing activity in the brain's reward circuitry (2011).

[55] L. L. Carstensen, D. M. Isaacowitz, S. T. Charles, Taking time seriously: a theory of socioemotional selectivity., American psychologist 54 (1999) 165.

[56] B. W. Roberts, K. E. Walton, W. Viechtbauer, Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies., Psychological bulletin 132 (2006) 1.

[57] T. O'Donoghue, M. Rabin, Doing It Now or Later, American Economic Review 89 (1999) 103–124. URL: https://pubs.aeaweb.org/doi/10.1257/aer.89.1.103. doi:10.1257/aer.89.1.103.

[58] C. Harris, D. Laibson, Dynamic Choices of Hyperbolic Consumers, Econometrica 69 (2001) 935–957. URL: http://doi.wiley.com/10.1111/1468-0262.00225. doi:10.1111/1468-0262.00225.

[59] G. Loewenstein, Out of control: Visceral influences on behavior, Organizational behavior and human decision processes 65 (1996) 272–292.

[60] G. F. Loewenstein, E. U. Weber, C. K. Hsee, N. Welch, Risk as feelings., Psychological bulletin 127 (2001) 267.

[61] D. Ariely, G. Loewenstein, The heat of the moment: The effect of sexual arousal on sexual decision making, Journal of behavioral decision making 19 (2006) 87–98.

[62] A. Mani, S. Mullainathan, E. Shafir, J. Zhao, Poverty impedes cognitive function, science 341 (2013) 976–980.

[63] B. Shiv, A. Fedorikhin, Heart and mind in conflict: The interplay of affect and cognition in consumer decision making, Journal of consumer Research 26 (1999) 278–292.

[64] J. Sweller, Cognitive load theory, in: Psychology of learning and motivation, volume 55, Elsevier, 2011, pp. 37–76.

[65] F. Lieder, T. L. Griffiths, Resource-rational analysis: Understanding human cognition as the

optimal use of limited computational resources, Behavioral and brain sciences 43 (2020) e1.

[66] J. W. Brehm, Postdecision changes in the desirability of alternatives., The Journal of Abnormal and Social Psychology 52 (1956) 384.

[67] L. Festinger, A theory of cognitive dissonance, A theory of cognitive dissonance, Stanford University Press, 1957. Pages: xi, 291.

[68] D. J. Bem, Self-perception theory, in: Advances in experimental social psychology, volume 6, Elsevier, 1972, pp. 1–62.

[69] P. Johansson, L. Hall, N. Chater, Preference Change through Choice, in: Neuroscience of Preference and Choice, Elsevier, 2012, pp. 121–141. URL: https://linkinghub.elsevier.com/retrieve/pii/B9780123814319000061. doi:10.1016/B978-0-12-381431-9.00006-1.

[70] V. Van Veen, M. K. Krug, J. W. Schooler, C. S. Carter, Neural activity predicts attitude change in cognitive dissonance, Nature Neuroscience 12 (2009) 1469–1474. URL: https://www.nature.com/articles/nn.2413. doi:10.1038/nn.2413.

[71] M. K. Chen, J. L. Risen, How choice affects and reflects preferences: revisiting the free-choice paradigm., Journal of personality and social psychology 99 (2010) 573.

[72] K. Izuma, K. Murayama, Choice-induced preference change in the free-choice paradigm: a critical methodological review, Frontiers in psychology 4 (2013) 41.

[73] D. Lee, J. Daunizeau, Choosing what we like vs liking what we choose: How choice-induced preference change might actually be instrumental to decision-making, PloS one 15 (2020) e0231081.

[74] T. Sharot, B. De Martino, R. J. Dolan, How choice reveals and shapes expected hedonic outcome, Journal of Neuroscience 29 (2009) 3760–3765.

[75] K. Voigt, C. Murawski, S. Speer, S. Bode, Hard decisions shape the neural coding of preferences, Journal of Neuroscience 39 (2019) 718–726.

[76] B. M. Staw, Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action, Organizational behavior and human performance 16 (1976) 27–44.

[77] H. R. Arkes, C. Blumer, The psychology of sunk cost, Organizational behavior and human decision processes 35 (1985) 124–140.

[78] S. Roth, T. Robbert, L. Straus, On the sunk-cost effect in economic decision-making: a meta-analytic review, Business research 8 (2015) 99–138.

[79] A. D. Martin, K. M. Quinn, Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999, Political analysis 10 (2002) 134–153.

[80] G. Loewenstein, Emotions in economic theory and economic behavior, American economic review 90 (2000) 426–432.

[81] J. D. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, Econometrica: Journal of the econometric society (1989) 357–384.

[82] B. Kőszegi, M. Rabin, A model of reference-dependent preferences, The Quarterly Journal of Economics 121 (2006) 1133–1165.

[83] J. Dewey, Human nature and conduct, volume 14, Southern Illinois University Press Carbondale, 1988.

[84] G. B. Saxe, Culture and cognitive development: Studies in mathematical understanding, Psychology Press, 2015.

[85] M. Heidegger, Basic writings: from Being and time (1927) to The task of thinking (1964) (1977).

[86] J. J. Gibson, The theory of affordances:(1979), in: The people, place, and space reader, Routledge, 2014, pp. 56–60.

[87] D. Norman, The design of everyday things: Revised and expanded edition, Basic books, 2013.

[88] M. McLuhan, Understanding media: The extensions of man, MIT press, 1994.

[89] H. H. Wilmer, L. E. Sherman, J. M. Chein, Smartphones and cognition: A review of research exploring the links between mobile technology habits and cognitive functioning, Frontiers in psychology 8 (2017) 605.

[90] N. Carr, The shallows: What the Internet is doing to our brains, WW Norton & Company, 2020.

[91] D. Kahneman, A. Tversky, Choices, values, and frames., American psychologist 39 (1984) 341.

[92] H. Schuman, S. Presser, J. Ludwig, Context effects on survey responses to questions about abortion, Public Opinion Quarterly 45 (1981) 216–223.

[93] R. Tourangeau, L. J. Rips, K. Rasinski, The psychology of survey response (2000).

[94] J. R. Busemeyer, P. D. Bruza, Quantum models of cognition and decision: principles and applications, second edition ed., Cambridge University Press, Cambridge, United Kingdom ; New York, NY, USA, 2025.

[95] J. R. Busemeyer, E. M. Pothos, R. Franco, J. S. Trueblood, A quantum theoretical explanation for probability judgment errors., Psychological review 118 (2011) 193.

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

[97] I. S. Maksymov, G. Pogrebna, The Physics of Preference: Unravelling Imprecision of Human Preferences through Magnetisation Dynamics, Information 15 (2024) 413. URL: https://www.mdpi.com/2078-2489/15/7/413. doi:10.3390/info15070413.

[98] G. De Tarde, The laws of imitation, H. Holt, 1903.

[99] M. S. Granovetter, The strength of weak ties, American journal of sociology 78 (1973) 1360–1380.

[100] H. Shirado, N. A. Christakis, Network engineering using autonomous agents increases cooperation in human groups, Iscience 23 (2020).

[101] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, Science 348 (2015) 1130–1132. URL: https://www.science.org/doi/10.1126/science.aaa1160. doi:10.1126/science.aaa1160.

[102] C. R. Sunstein, #Republic: divided democracy in the age of social media, paperback edition ed., Princeton University Press, Princeton, 2018.

[103] A. Pentland, Social physics: How good ideas spread-the lessons from a new science, Penguin, 2014.

[104] F. Huszár, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, M. Hardt, Algorithmic amplification of politics on Twitter, Proceedings of the national academy of sciences 119 (2022) e2025334119.

[105] A. D. I. Kramer, J. E. Guillory, J. T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, Proceedings of the National Academy of Sciences 111 (2014) 8788–8790. URL: https://www.pnas.org/doi/10.1073/pnas.1320040111. doi:10.1073/pnas.1320040111.

[106] H. Allcott, L. Braghieri, S. Eichmeyer, M. Gentzkow, The Welfare Effects of Social Media, American Economic Review 110 (2020) 629–676. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20190658. doi:10.1257/aer.20190658.

[107] A. J. Chaney, B. M. Stewart, B. E. Engelhardt, How algorithmic confounding in recommendation systems increases homogeneity and decreases utility, 2018, pp. 224–232.

[108] Y. Engeström, Expansive Learning at Work: Toward an activity theoretical reconceptualization, Journal of Education and Work 14 (2001) 133–156. URL: http://www.tandfonline.com/doi/abs/10.1080/13639080020028747. doi:10.1080/13639080020028747.

[109] Y. Engeström, A. Sannino, From mediated actions to heterogenous coalitions: four generations of activity-theoretical studies of work and learning, Mind, culture, and activity 28 (2021) 4–23.

[110] M. Aristotle, The Nicomachean Ethics.(D. Ross, Trans.) (1998).

[111] L. Kohlberg, The philosophy of moral development: Moral stages and the idea of justice (1981).

[112] E. Turiel, The development of social knowledge: Morality and convention, Cambridge University Press, 1983.

[113] J. R. Rest, Moral development: Advances in research and theory (1986).

[114] A. Gutmann, Democratic education (1987).

[115] S. J. Khader, Adaptive preferences and women's empowerment, Oxford University Press, 2011.

[116] D. Friedman, C. Diem, Rational-Choice Theory, Feminist Critiques, and Gender Inequality, Theory on gender: Feminism on theory (1993) 91.

[117] I. M. Young, Justice and the Politics of Difference, Princeton university press, 1990.

[118] N. Volkow, M. Morales, The Brain on Drugs: From Reward to Addiction, Cell 162 (2015) 712–725. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867415009629. doi:10.1016/j.cell.2015.07.046.

[119] L. O. Gostin, Public health law: power, duty, restraint, volume 3, Univ of California Press, 2000.

[120] T. L. Beauchamp, J. F. Childress, Principles of biomedical ethics, Edicoes Loyola, 1994.

[121] R. H. Thaler, C. R. Sunstein, Nudge: improving decisions about health, wealth and happiness, revised edition, new international edition ed., Penguin Books, London New York Toronto Dublin Camberwell New Delhi Rosedale Johannesburg, 2009.

[122] I. Gabriel, Artificial intelligence, values, and alignment, Minds and machines 30 (2020) 411–437.

[123] A. Y. Ng, S. Russell, Algorithms for inverse reinforcement learning., volume 1, 2000, p. 2.

[124] D. Hadfield-Menell, S. J. Russell, P. Abbeel, A. Dragan, Cooperative inverse reinforcement learning, Advances in neural information processing systems 29 (2016).

[125] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in neural information processing systems 30 (2017).

[126] S. Russell, Human compatible: AI and the problem of control, Penguin Uk, 2019.

[127] A. R. Doshi, O. P. Hauser, Generative AI enhances individual creativity but reduces the collective diversity of novel content, Science Advances 10 (2024) eadn5290. URL: https://www.science.org/doi/10.1126/sciadv.adn5290. doi:10.1126/sciadv.adn5290.

[128] P. Abbeel, A. Y. Ng, Apprenticeship learning via inverse reinforcement learning, 2004, p. 1.

[129] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving, Fine-Tuning Language Models from Human Preferences, 2020. URL: http://arxiv.org/abs/1909.08593. doi:10.48550/arXiv.1909.08593, arXiv:1909.08593 [cs].

[130] D. Krasheninnikov, R. Shah, H. v. Hoof, Combining Reward Information from Multiple Sources, 2021. URL: http://arxiv.org/abs/2103.12142. doi:10.48550/arXiv.2103.12142, arXiv:2103.12142 [cs].

[131] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. URL: http://arxiv.org/abs/2203.02155. doi:10.48550/arXiv.2203.02155, arXiv:2203.02155 [cs].

[132] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in AI safety, arXiv preprint arXiv:1606.06565 (2016).

[133] N. Soares, B. Fallenstein, S. Armstrong, E. Yudkowsky, Corrigibility., 2015.

[134] P. Christiano, B. Shlegeris, D. Amodei, Supervising strong learners by amplifying weak experts, arXiv preprint arXiv:1810.08575 (2018).

[135] G. Irving, P. Christiano, D. Amodei, AI safety via debate, arXiv preprint arXiv:1805.00899 (2018).

[136] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, S. Legg, Scalable agent alignment via reward modeling: a research direction, arXiv preprint arXiv:1811.07871 (2018).

[137] E. Yudkowsky, Coherent extrapolated volition, Singularity Institute for Artificial Intelligence (2004).

[138] O. Evans, N. D. Goodman, Learning the preferences of bounded agents, volume 6, 2015, pp. 2–1.

[139] R. Shah, N. Gundotra, P. Abbeel, A. D. Dragan, On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference, 2019. URL: http://arxiv.org/abs/1906.09624. doi:10.48550/arXiv.1906.09624, arXiv:1906.09624 [cs].

[140] A. Fickinger, S. Zhuang, D. Hadfield-Menell, S. Russell, Multi-principal assistance games, arXiv preprint arXiv:2007.09540 (2020).

[141] A. Kierans, A. Ghosh, H. Hazan, S. Dori-Hacohen, Quantifying misalignment between agents: Towards a sociotechnical understanding of alignment, volume 39, 2025, pp. 27365–27373.

[142] V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf, E. Tewolde, W. S. Zwicker, Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback, 2024. URL: http://arxiv.org/abs/2404.10271. doi:10.48550/arXiv.2404.10271, arXiv:2404.10271 [cs].

[143] T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, A roadmap to pluralistic alignment, arXiv preprint arXiv:2402.05070 (2024).

[144] S. Zhao, J. Dang, A. Grover, Group Preference Optimization: Few-Shot Alignment of Large Language Models, 2024. URL: http://arxiv.org/abs/2310.11523. doi:10.48550/arXiv.2310.11523, arXiv:2310.11523 [cs].

[145] M. Srewa, T. Zhao, S. Elmalaki, PluralLLM: Pluralistic Alignment in LLMs via Federated Learning, in: Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems, HumanSys '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 64–69. URL: https://dl.acm.org/doi/10.1145/3722570.3726898. doi:10.1145/3722570.3726898.

[146] M. Srewa, T. Zhao, S. Elmalaki, A Systematic Evaluation of Preference Aggregation in Federated RLHF for Pluralistic Alignment of LLMs, 2025. URL: https://openreview.net/forum?id=vfP16cLfH0.

[147] C. J. Bates, R. Bose, R. G. Keeney, V. A. Kazakova, Contractual AI: Toward More Aligned, Transparent, and Robust Dialogue Agents, volume 2, 2023, pp. 225–227.

[148] S. Levine, M. Franklin, T. Zhi-Xuan, S. Y. Guyot, L. Wong, D. Kilov, Y. Choi, J. B. Tenenbaum, N. Goodman, S. Lazar, Resource Rational Contractualism Should Guide AI Alignment, arXiv preprint arXiv:2506.17434 (2025).

[149] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, Constitutional AI: Harmlessness from AI Feedback, 2022. URL: http://arxiv.org/abs/2212.08073. doi:10.48550/arXiv.2212.08073, arXiv:2212.08073 [cs].

[150] S. Huang, D. Siddarth, L. Lovitt, T. I. Liao, E. Durmus, A. Tamkin, D. Ganguli, Collective constitutional ai: Aligning a language model with public input, 2024, pp. 1395–1417.

[151] S. Lazar, A. Nelson, AI safety on whose terms?, Science 381 (2023) 138–138.

[152] J. Kroll, J. Huey, S. Barocas, E. Felten, J. Reidenberg, D. Robinson, H. Yu, Accountable Algorithms, University of Pennsylvania Law Review 165 (2017) 633. URL: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

[153] J. Edelman, T. Zhi-Xuan, R. Lowe, O. Klingefjord, V. Wang-Mascianica, M. Franklin, R. O. Kearns, E. Hain, A. Sarkar, M. Bakker, F. Barez, D. Duvenaud, J. Foerster, I. Gabriel, J. Gubbels, B. Goodman, A. Haupt, J. Heitzig, J. Jara-Ettinger, A. Kasirzadeh, J. R. Kirkpatrick, A. Koh, W. B. Knox, P. Koralus, J. Lehman, S. Levine, S. Marro, M. Revel, T. Shorin, M. Sutherland, M. H. Tessler, I. Vendrov, J. Wilken-Smith, Full-Stack Alignment: Co-Aligning AI and Institutions with Thick Models of Value, 2025. URL: http://arxiv.org/abs/2512.03399. doi:10.48550/arXiv.2512.03399, arXiv:2512.03399 [cs].

[154] J. R. Anthis, D. Asmar, K. R. Driggs-Campbell, A. Hardy, K. J. Meimandi, G. Keeling, M. Kochenderfer, H. Liu, S. Liu, R. Martín-Martín, A. A. Rushdi, M. R. Schlichting, P. Stone, H. Subramonyam, D. Yang, ICLR 2025 Workshop on Human-AI Coevolution, 2024. URL: https://openreview.net/forum?id=0K45ILUlYM.

[155] J. Li, T. Song, B. Xue, Y.-C. Lee, We Shape AI, and Thereafter AI Shape Us: Humans Align with AI through Social Influences, 2025. URL: https://openreview.net/forum?id=64rCWVC78p.

[156] C. Tanguy, R. Janssens, T. Belpaeme, J. Dambre, Human Alignment: How Much Do We Adapt to LLMs?, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 603–613. URL: https://aclanthology.org/2025.acl-short.47. doi:10.18653/v1/2025.acl-short.47.

[157] J. Li, Y. Yang, Q. V. Liao, J. Zhang, Y.-C. Lee, As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making, in: Proceedings of

the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–16. URL: https://dl.acm.org/doi/10.1145/3706598.3713336. doi:10.1145/3706598.3713336.

[158] R. Mushkani, H. Berard, S. Koseki, Negotiative Alignment: Embracing Disagreement to Achieve Fairer Outcomes – Insights from Urban Studies, 2025. URL: http://arxiv.org/abs/2503.12613. doi:10.48550/arXiv.2503.12613, arXiv:2503.12613 [cs].

[159] M. Carroll, A. Foote, K. Feng, M. Williams, A. Dragan, W. B. Knox, S. Milli, CTRL-Rec: Controlling Recommender Systems With Natural Language, 2025. URL: http://arxiv.org/abs/2510.12742. doi:10.48550/arXiv.2510.12742, arXiv:2510.12742 [cs].

[160] M. Carroll, D. Foote, A. Siththaranjan, S. Russell, A. Dragan, Ai alignment with changing and influenceable reward functions, arXiv preprint arXiv:2405.17713 (2024).

[161] C. Tran, K. Fasiang, M. Kanwal, E. O'Rourke, Starting from scratch again and again: Tracing the origins of high schoolers' negative perceptions of block-based programming, in: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), Association for Computing Machinery, 2026. To appear.

[162] C. Salge, C. Glackin, D. Polani, Empowerment – an introduction, 2013. URL: https://arxiv.org/abs/1310.1863. arXiv:1310.1863.