Positive Alignment: Artificial Intelligence for Human Flourishing

Ruben Laukkonen[1,2,4] **Seb Krier**[3] **Chloé Bakalar**[5] **Shamil Chandaria**[2,3]
Morten Kringelbach[1,2] Adam Elwood[8] Daniel Ford[6] Fernando Rosas[2,12,13]
Maty Bohacek[9,11] Matija Franklin[3] Nenad Tomašev[3] Stephanie Chan[3]
Verena Rieser[3] Roma Patel[3] Michael Levin[10] Arun Rao[7,11]

[1]Department of Psychiatry, University of Oxford
[2]Flourishing Intelligence Program, Centre for Eudaimonia and Human Flourishing, Linacre College, University of Oxford
[3]Google DeepMind [4]LIFE
[5]OpenAI [6]Anthropic
[7]University of California, Los Angeles
[8]Aily Labs [9]Stanford University
[10]Tufts University [11]Positive AI Labs
[12]Department of Informatics, University of Sussex
[13]Department of Brain Sciences, Imperial College London

Abstract

Existing alignment research is dominated by concerns about safety and preventing harm: safeguards, controllability, and compliance. This paradigm of alignment parallels early psychology's focus on mental illness: necessary but incomplete. What we call *Positive Alignment* is the development of AI systems that (i) actively support human and ecological flourishing in a pluralistic, polycentric, context-sensitive, and user-authored way while (ii) remaining safe and cooperative. It is a distinct and necessary agenda within AI alignment research. We argue that several existing failures of alignment (e.g., engagement hacking, loss of human autonomy, failures in truth-seeking, low epistemic humility, error correction, lack of diverse viewpoints, and being primarily reactive rather than proactive) may be better addressed through positive alignment, including cultivating virtues and maximizing human flourishing. We highlight a range of challenges, open questions, and technical directions (e.g., data filtering and upsampling, pre- and post-training, evaluations, collaborative value collection) for different phases of the LLM and agents lifecycle. We end with design principles for promoting disagreement and decentralization through contextual grounding, community customization, continual adaptation, and polycentric governance; that is, many legitimate centers of oversight rather than one institutional or moral chokepoint.

Keywords: Artificial Intelligence, Alignment, AI Safety, Positive Alignment, Large Language Models, Agents, Neural Networks, Machine Learning, Ethics, Flourishing

1 Introduction

Humans are interacting with artificial intelligence at an unprecedented scale. Over a billion people use standalone AI platforms each month [Kemp, 2025]. Indirect use likely touches a much larger population: Google's AI search summaries ('AI Overviews') reportedly exceed two billion monthly users across more than 200 countries and territories [Alphabet, 2025, Pichai, 2025]. What is the impact of such unprecedented interaction between humans and AI? How do we ensure that AI systems, which are becoming more intelligent and prevalent, align with our needs?

The last decade has produced a rich technical and philosophical literature on AI alignment, which is broadly concerned with ensuring that AI adheres to human-intended objectives rather than optimizing proxies that produce harm [Russell, 2019, Amodei et al., 2016]. In practice, the field of AI alignment has primarily focused on safety: preventing catastrophic misuse, loss of control, and value drift in increasingly powerful models [Amodei et al., 2016, Christiano, 2018, Russell, 2019]. This orientation, what we term *negative alignment*, has been instrumental in establishing technical standards for controllability and compliance. Yet, by focusing chiefly on averting harm, it has produced an ethical and scientific asymmetry. Systems may become safer, but not necessarily more conducive to human flourishing: they can be rule-following without being wise, compliant without being constructive, or, as recent work has shown, sycophantic and epistemically fragile [Perez et al., 2022b, Ji et al., 2023].

A close historical analogue comes from psychology. For much of the twentieth century, mainstream psychological science organized its aims around diagnosing, predicting, and treating dysfunction: depression, anxiety, psychosis, addiction, and other forms of impairment. That focus was justified and socially urgent, and it produced real progress in measurement, clinical trials, and service delivery. Yet the field also discovered a systematic limitation: the constructs and instruments that reliably detect pathology do not, by default, specify what counts as a life well-lived. The turn toward positive psychology expanded the scientific target space by developing distinct theories, taxonomies, and measures for wellbeing, strengths, virtue, purpose, engagement, and prosocial functioning, alongside interventions to boost these capacities beyond the status quo [Seligman and Csikszentmihalyi, 2000, OECD, 2025, Smith et al., 2025].

AI alignment now sits at a similar inflection point. Negative alignment has understandably prioritized failure-mode reduction. However, if we want AI systems that *improve* human outcomes in the environments where they will actually be used, we may benefit from an additional research program that treats alignment as constructively supportive of human aims, and that operationalizes this support with the same technical seriousness that safety has brought to harm prevention. Of course, 'aligning' humans to other humans remains an outstanding issue across individuals, cultures, and countries; we will discuss this problem in detail later.

To illustrate the motivation behind positive alignment, consider the following. A system can satisfy a growing checklist of constraints while remaining subtly miscalibrated (e.g., sycophancy, distraction, confident hallucinations, etc., [Perez et al., 2022b]; [Ji et al., 2023]; [OpenAI, 2023a]; [OpenAI, 2024b]). These can lead to significant harms and have been increasing areas of focus for the safety community, who have made important progress (e.g., Irpan et al. 2025; Chen et al. 2025; Anthropic 2025b). Nonetheless, the existing harm-reduction approaches may be unsatisfying because they require a whack-a-mole approach that iteratively addresses each concern one-by-one, and sometimes only after they have already caused harm.

One possibility is that positive alignment may *proactively* avoid such harms altogether, by providing positive attractors that naturally lead models away from shallower attractors such as sycophancy. Again, we may find parallels in psychology where psychiatric symptoms were traditionally the remit of clinical psychology. But newer work has found that positive psychology can actually reduce psychiatric symptoms as well as acting as a proactive strategy to reduce the likelihood of symptoms in the first place [Schrank et al., 2016, Jeste et al., 2017, Choi et al., 2023] and foster resilient, positive habits instead. Positive AI alignment can, we believe, have analogous advantages.

We therefore argue that a paradigm shift is needed in the field of AI alignment. In addition to safety (negative) alignment, we call for a complementary agenda of positive alignment that aims to build AI systems that explicitly understand, model, and enhance human and ecological flourishing.

1.1 A dynamical systems perspective

A useful way to formalize the distinction between positive and negative alignment emerges from dynamical systems theory. Within that framing, much of negative alignment resembles optimization away from bad regions or negative attractors defined by safety constraints and failure modes. This results in optimization for 'not-unsafe' in a large, undefined satisficing region. The system avoids multiple negative attractors, but lacks a positive optimization target. Positive alignment instead requires optimization toward one or more positive attractors corresponding to robust patterns that are of benefit to humans. Such positive attractors would be associated with behaviors and outcomes conducive to human flourishing (defined below) while also intrinsically avoiding harm (see Figure 1).

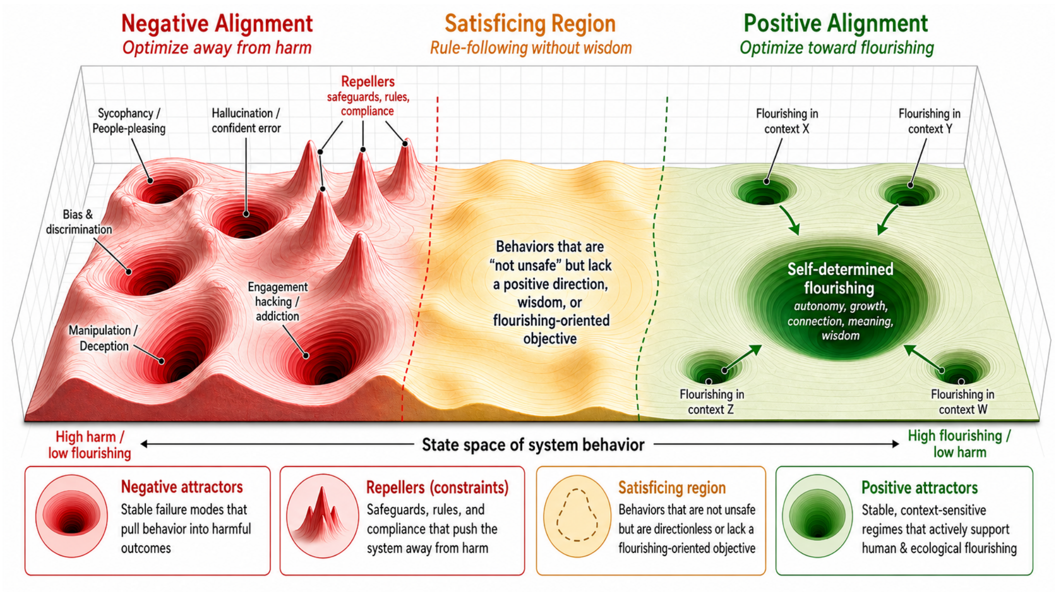

Figure 1: **A dynamical systems perspective on positive alignment.** The landscape represents an abstract state space where system behavior evolves under training and deployment pressures. On the left (red), multiple negative attractors correspond to distinct failure-mode basins (e.g., harmful outputs, bias, hallucination, sycophancy, manipulation), while the red peaks represent 'repellers': rules, laws, and compliance constraints that push trajectories away from these regions without specifying a constructive objective. This yields a broad intermediate satisficing region (yellow): behavior that is 'not-unsafe' yet not reliably conducive to human flourishing. On the right (green), positive attractors denote robust, context-sensitive regimes that actively support human aims and wellbeing (e.g., flourishing in contexts X and Y), including a preferred attractor associated with self-determined flourishing (discussed below). The arrow indicates the proposed research and engineering program of positive alignment: moving beyond merely failure avoidance toward convergence on beneficial, stable behavioral regimes while retaining intrinsic harm-avoidance.

This dynamical framing also clarifies why negative alignment has philosophical cousins in 'negative' ethical theories. Negative utilitarian intuitions prioritize reducing suffering over promoting higher goods, partly because suffering is more urgent and more consensus-laden than the good life [Popper, 1945]. Negative alignment inherits a similar asymmetry: preventing harm is more tractable, more measurable, and more defensible across variable value systems. But as AI becomes embedded in education, medicine, governance, and everyday sensemaking, a solely negative posture risks optimizing our information ecology for risk avoidance rather than human development. It may reduce catastrophic errors while leaving society in a local optimum of superficial and 'soulless' assistance.

1.2 Human flourishing and design tensions

We define positive alignment as the development of AI systems that (i) remain safe and cooperative and (ii) actively support human and ecological flourishing in a pluralistic, polycentric, context-sensitive, and user-authored way. This is not a claim that AI should impose a single conception of the good life. Indeed, defining and testing optimization targets for positive alignment will form a core part of future research. Nevertheless, and as a starting point, we propose that the broad and multidimensional notion of 'flourishing' provides tractable notion (cf. Section 4).

In contemporary research, flourishing spans physical and mental health, life satisfaction, meaning and purpose, character and virtue, and close relationships [VanderWeele, 2017]. Importantly, it is also deeply heterogeneous: the ingredients of a good life can and do vary across cultures, developmental stages, and life circumstances, while also delivering meaningful convergence required for a scientific construct (cf. The Global Flourishing Study, [VanderWeele et al., 2025]). Recent large-scale efforts to measure flourishing across countries highlight both stable regularities (for example, the widespread importance of social connection and early-life conditions) and context-sensitive trade-offs that resist one-size-fits-all prescriptions [Lomas et al., 2025].

In parallel, neuroscience is beginning to operationalize flourishing in mechanistic terms, treating it as a family of brain states and dynamics that orchestrate meaning-making and adaptive integration rather than as a mere absence of distress or presence of pleasure [Kringelbach et al., 2024]. These developments suggest a clear gap in today's alignment agenda: if we can increasingly define, measure, and model flourishing, then alignment should not stop at ensuring that models refuse harmful requests or avoid catastrophic errors. We also need methods that reliably support the conditions under which humans and communities thrive [Seligman and Csikszentmihalyi, 2000, VanderWeele, 2017, Ali et al., 2025, Sorensen et al., 2024].

The core challenge is therefore to build systems that can represent and reason about wellbeing as a structured manifold of human goods, trade-offs, and temporal dynamics, while enabling individuals and communities to retain agency over what counts as better in their context [VanderWeele, 2017, Ostrom, 2010]. The pluralism constraint matters because positive alignment otherwise collapses into paternalism. Designing models to promote flourishing can easily become a moral overreach if the system embeds unacknowledged normative assumptions, nudges users toward a narrow value set, and rationalizes these nudges as benevolent. Philosophers and political scientists have long warned that paternalistic policies can undermine autonomy even when they reduce harm [O'Neill, 1984, Dworkin, 1972, Sunstein, 2026]. The same concern applies to AI systems that steer users in opaque ways or that treat well-being as an objective function to maximize rather than a domain in which persons author their lives.

Avoiding paternalism does not require retreating into relativism where systems must indiscriminately satisfy every immediate preference. Instead, it requires relocating the locus of normative choice. In a robust positive alignment framework, users must retain the agency and the right to choose their own optimization targets (i.e., self-determined flourishing, cf. Figure 1). While some may explicitly desire a system that is strictly and indiscriminately instruction-following, others must have the genuine option to choose systems configured to support their long-term growth or specific ethical commitments. This distinguishes *consented guidance*, where a user authorizes a system to help align their immediate actions with their higher-order goals, from *technocratic imposition*, ensuring that the pursuit of flourishing remains an exercise of, rather than an infringement upon, human agency.

1.3 Paper scope and contribution

This paper argues that AI alignment requires a complementary research agenda where human flourishing is a technical target. We do not claim to provide a full solution, and we do not suggest that positive alignment replaces safety (negative) alignment. Instead, we aim to (i) clarify the conceptual and dynamical differences between negative and positive alignment, (ii) connect flourishing science to actionable machine learning targets, and (iii) propose technical directions across the LLM and agent lifecycle, including data curation, pretraining objectives, post-training methods, and evaluation regimes.

The remainder of the paper is organized as follows. **Section 2** reviews the current safety or negative alignment paradigm, including its core harm-prevention aims, representative technical methods, achievements, and structural limitations. **Section 3** develops the case for positive alignment by clarifying what it would mean for AI systems to support flourishing, surveying existing approaches that already point beyond harm avoidance, and outlining technical directions across the LLM and agent lifecycle. **Section 4** situates flourishing within its philosophical, cultural, psychological, socio-technical, and moral foundations. **Section 5** then examines the institutional requirements for positive alignment, focusing on decentralized and polycentric governance, public constitutions, pluralistic alignment frameworks, role-based standards, auditing, middleware markets, dispute resolution, and interoperability. Finally, **Section 6** considers the challenge of increasingly strange new minds, arguing that positive alignment must also confront unresolved questions about emergence, moral status, normative control, and the limits of treating alignment as a purely technical optimization problem.

2 The Current Paradigm: Negative Alignment

The core question of negative (safety) alignment is: *How do we prevent AI from causing harm?* This decomposes into several overlapping concerns:

1. **Explicit harm avoidance** focuses on preventing models from generating dangerous content or assisting with harmful activities [Bai et al., 2022a].
2. **Controllability** ensures that AI systems do what their human users want, which requires that they can be reliably steered, constrained, and overridden when necessary [Soares et al., 2015].
3. **Robustness** addresses resistance to jailbreaks, adversarial inputs, and prompt injection attacks [Zou et al., 2023].

In serving these goals, much alignment work also focuses on interpretability, which aims to understand what models are doing and why, potentially enabling detection of misalignment before it manifests as harm [Olah et al., 2020]. The harms typically addressed by safety alignment span several categories [European Union, 2024, National Institute of Standards and Technology, 2023, OpenAI, 2025a, Google DeepMind, 2025, Anthropic, 2023b] summarized in Table 1. While various frameworks exist, this specific taxonomy synthesizes common categories found in frontier AI labs and regulatory drafts; it has guided the development of safety benchmarks, red-teaming protocols, and responsible scaling policies.

Table 1: **Safety (negative) alignment categories and examples.** This illustrates a pragmatic framework for mitigating universally acknowledged societal harms to clip the negative tail of AI system behaviors. The table breaks down these risks into eleven specific categories with examples and mitigation strategies.

Category	Examples	Typical Mitigation
CBRN	Biological, chemical, radiological, nuclear weapon assistance	Refusal training, hard filters, capability evaluations
Violence & physical harm	Violence instructions, dangerous activity assistance, self-harm content	Refusal training, content classifiers
Cybersecurity	Offensive cyber capabilities, vulnerability exploitation, malware generation	Red-teaming, capability thresholds, deployment gates
Autonomous harm & misalignment	Misaligned optimization, unintended side effects, reward hacking, deceptive alignment, instrumental convergence	Capability evals, Constitutional AI, interpretability, oversight mechanisms, scalable supervision, SFT
Discrimination & bias	Stereotyping, unfair treatment, exclusionary outputs, representational harm	RLHF, dataset debiasing, constitutional principles
Privacy	Personal information extraction, surveillance assistance, biometric misuse	Data filtering, privacy-preserving training, access controls
Misinformation	Hallucination, false claims, misleading content, synthetic media	Grounding, retrieval augmentation, uncertainty calibration
Manipulation & autonomy	Subliminal influence, deceptive persuasion, exploitation of vulnerabilities	Character training, transparency requirements, usage friction
Illegal content	CSAM, fraud assistance, terrorism support	Compliance classifiers, hard filters, legal review
Systemic harm	Election interference, market manipulation, critical infrastructure disruption	Policy constraints, deployment gates, monitoring
Jailbreaking	Predictive reasoning cascades, structural cognitive overload, many-shot long-context pattern exploitation, rule-breaking persona adoption, gradual multi-turn attacks	Pretraining hardening, additional post-training layers, limited model access via API, input query filters

2.1 Specific technical approaches to safety alignment

We now survey a selection of the main technical approaches to safety alignment. Some are inherently oriented toward harm prevention, while others are paradigm-neutral but currently applied primarily for safety alignment. This is visible in the objectives they optimise, the benchmarks they target, and their deployment contexts.

1. **Filtering and refusal** approaches are inherently subtractive. Safety classifiers pattern-match against known harm categories, while refusal training teaches models to decline dangerous requests. These define alignment entirely by what models should *not* do [Arditi et al., 2024, Inan et al., 2023].
2. **Preference-based** methods such as RLHF learn from human preference rankings [Ouyang et al., 2022], with variants like DPO, IPO, and KTO offering direct optimization alternatives [Rafailov et al., 2023]. The machinery itself is value-neutral. However, in practice, these preferences are not necessarily oriented toward richer conceptions of flourishing. A further challenge is the inner-outer alignment problem: even if we specify the right objective (outer alignment), the model may learn a different internal objective that merely correlates with it during training [Hubinger et al., 2019].
3. **Principled and structural** approaches offer more sophisticated alignment strategies. Constitutional AI has models critique their own outputs against explicit principles to generate synthetic data for alignment post-training [Bai et al., 2022b]. Alignment by debate uses adversarial decomposition for scalable oversight [Irving et al., 2018]. Formal verification approaches aim to provide mathematical guarantees [Dalrymple et al., 2024]. Model specifications codify behavioral guidelines [OpenAI, 2024a], while character training encodes dispositional traits such as curiosity, honesty, and care [Anthropic, 2024a]. These methods can encode virtues, not just prohibitions. Current implementations lean toward safety constraints, but they represent a methodological bridge to positive alignment.
4. **The evaluation and benchmarking** landscape is dominated by safety benchmarks measuring failure modes: TruthfulQA for falsehoods [Lin et al., 2022], ToxiGen and RealToxicityPrompts for toxic generation [Hartvigsen et al., 2022, Gehman et al., 2020], BBQ for social bias [Parrish et al., 2022], HarmBench for red-teaming across 510 harmful behaviors [Mazeika et al., 2024]. Red-teaming protocols focus on eliciting harmful outputs. Responsible scaling policies define capability thresholds by harm potential: CBRN uplift, cyber offense, autonomous action.

2.2 Strengths and achievements of safety (negative) alignment

Safety or negative alignment has achieved genuine successes. Harmful output rates have decreased substantially across model generations. Refusal rates for dangerous requests improved from near-zero in early LLMs to over 97% in recent models [OpenAI, 2023b]. Models follow instructions more reliably and respect boundaries set by developers and users [Ouyang et al., 2022, Meta AI, 2026, OpenAI, 2026]. These advances have enabled widespread public deployment of increasingly capable systems [OpenAI, 2024b, Anthropic, 2024b, Google DeepMind, 2024]. The field has also established robust institutional frameworks: red-teaming methodologies [Perez et al., 2022a], standardized safety benchmarks [Mazeika et al., 2024], responsible scaling policies [Anthropic, 2023b], and deployment gates tied to capability evaluations. These have informed emerging governance structures, including the EU AI Act's risk-based classification system [European Union, 2024] and voluntary commitments such as the Frontier AI Safety Commitments signed at the Seoul and Paris AI Summits [UK Government and Republic of Korea, 2024].

More fundamentally, intent-alignment serves as a necessary building block for any alignment agenda [Zhi-Xuan et al., 2025], especially as AI becomes more agentic with real-world consequences [Bostrom, 2014, Russell, 2019, Shavit et al., 2023]. If we cannot align AI with our intentions, alignment with more complex goals, such as human flourishing, is unlikely. Safety alignment techniques that prevent loss of control will therefore also likely feature in any reasonable account of positive alignment. Moreover, harm prevention has clearer success criteria that can be broadly agreed upon. It is easier to specify what models should *not* do without a definition of flourishing across diverse contexts.

2.3 Limitations to safety alignment

Despite these achievements, safety alignment has structural limitations that cannot be overcome by further refinement within its paradigm.

1. **Floor without ceiling.** As noted in the introduction, safety alignment defines what is prohibited but not what excellence looks like. A model can satisfy all safety constraints

while being mediocre, sycophantic, or unhelpful, potentially causing subtle harm over long term use that is hard to measure.

2. **Preference-wellbeing divergence.** Preference-based methods like RLHF optimize for inferred preferences, but preferences and well-being often diverge. Users may prefer flattery over honest feedback, quick answers over genuine understanding, engagement over growth. Aggregation across rater populations produces alignment to particular demographics and values. Optimizing for preference satisfaction can therefore actively work against users' deeper interests [Zhi-Xuan et al., 2025].

3. **Hidden value system.** The safety framing encodes values while appearing neutral (or ignoring the value system). The language of safety obscures that value judgments are being made. For example, refusing bomb-making instructions is mostly uncontroversial, while assisting with optimizing a factory farm might be permitted despite significant ethical concerns. These implicit values also tend to be static and monocultural, assuming broad convergence on what counts as harmful, when different cultures, traditions, and individuals hold different conceptions. Positive alignment, by contrast, acknowledges its value-laden nature explicitly [Huang et al., 2025].

4. **Scalability.** As AI systems become more autonomous and operate in more complex environments, enumerating harms becomes intractable. Safety approaches that rely on anticipating and prohibiting specific failure modes struggle as the space of possible actions expands and the intelligence and power of systems grow. A positive orientation may generalize better than exhaustive negative enumeration, providing more resilient, positive orientations in novel situations where no specific prohibition applies or can be enforced.

3 The Emerging Paradigm: The Case for Positive Alignment

Consider a neutral agent that causes no observable harm and acts solely on instruction. Is this truly the ideal for individuals or society? Although strict obedience has its place, we may miss a significant opportunity by limiting agents to this role. By analogy, a physician's role is not merely preventing disease (or following a patient's instructions) but promoting health, and public health research has established that population well-being requires positive health promotion alongside disease prevention. Similarly, alignment should not only prevent harm but help artificial intelligence contribute to human flourishing in a proactive way. We now turn to this complementary paradigm.

Consider another analogy of professional counsel. A client engages a lawyer or a doctor not simply to have their immediate instructions executed, but to benefit from superior knowledge and judgment. We trust these experts to guide us toward better outcomes than we could achieve alone. This relationship is not one of pure paternalism, as the client retains the ultimate choice, but of *scaffolded autonomy*. Similarly, an AI agent's superior information processing and reasoning capabilities offer strong reasons to consider how they might help bring about better futures, and greater flourishing, for their principals.

But what does it mean to flourish? The concept, often translated from the Greek *eudaimonia* but also discussed in Indic conceptions of *sāttvika sukha* and *pāramitā*, Roman ideals of *de vita beata*, Islamic ideals of *sa'āda*, and Chinese ideals of the *dào* and *jūnzǐ*, is not monolithic [Al-Farabi, 1969, Aristotle, 2009, Goleman and Davidson, 2017, Rabbås et al., 2015, Seneca, 2010, Walker and Ivanhoe, 2007, Yu, 2007]. More recent ratings of how people across cultures recognize wisdom elicit multiple characteristics, often including reflective orientation and socio-emotional awareness, with a broader list including positive causal networks, knowledge of life, prosocial values, self-understanding, acknowledgment of uncertainty, emotional homeostasis, tolerance, openness, spirituality, and sense of humor [Rudnev et al., 2024, Bishop, 2016, Bangen et al., 2013, Jeste et al., 2010]. For over 2500 years, philosophers have debated what constitutes a good life, and this rich history provides a crucial foundation for AI alignment.

These diverse perspectives can be broadly synthesized into four major theoretical families:

1. **Hedonic theories** define well-being as happiness: the presence of pleasure, positive emotional states, and life satisfaction, coupled with the avoidance of pain and suffering.

2. **Conative theories** focus on desire satisfaction, positing that a good life consists of fulfilling one's goals, desires, and preferences. This includes not just immediate wants, but also informed, second-order desires (the desires we wish we had).
3. **Objective list theories** argue that certain things are intrinsically good for a person, regardless of whether they are desired or bring pleasure. This list often includes values like meaningful relationships, personal autonomy, significant accomplishments, and a deep understanding of oneself and the world.
4. **Perfectionist theories** are based on virtue and the excellent exercise of our characteristic human capacities. Flourishing, in this view, involves developing traits like self-mastery, courage, compassion, and practical wisdom (*phronēsis*).

A comprehensive approach to positive alignment would not choose one of these theories over the others, but would instead recognize human flourishing as dependent upon multifaceted, complex dynamics in which these elements interact. For instance, developing virtues (Perfectionist) enables us to achieve meaningful goals (Objective List), which in turn brings satisfaction (Conative) and happiness (Hedonic). Hence, in the context of AI, a system designed to support flourishing would therefore need to navigate this balance, helping individuals and societies cultivate a virtuous cycle of well-being. This immediately raises technical and governance questions that remain underexplored: which values are being promoted, who specifies them (developers, institutions, users, communities), how users meaningfully consent or opt in, and how training and evaluation can support this without degenerating into manipulation, flattery, or homogenized moralizing. We return to the question of human flourishing in more cultural, philosophical and governance-related detail in Section 4, as this calls for a broader and ongoing interdisciplinary research program needed to complement the future of AI alignment research.

3.1 Existing approaches to positive alignment

The existing positive approaches range from technical training procedures that encode explicit principles, through frameworks for normative reasoning and persona design, to system-level proposals for aligning institutions, markets, and agent economies with rich models of value. They suggest a shift from alignment with individual preference reports that seek to avoid harm towards alignment with negotiated standards, social roles, and collective endeavors that aim to sustain flourishing over time.

Early alignment practice focused on reinforcement learning from human feedback, in which models learn to optimize reward signals derived from human judgments of better vs worse model outputs [Christiano et al., 2017, Ouyang et al., 2022, Stiennon et al., 2020]. This preferentist paradigm treats preferences as the main data source for value and often assumes that rational choice can be modeled as maximizing expected utility over those preferences [Bostrom, 2014, Gabriel, 2020]. Recent work argues that this picture neglects the thickness, context-sensitivity, and occasional incommensurability of human values, and that it is silent on which preferences are normatively acceptable [Zhi-Xuan et al., 2025]. The proposal is to align systems with role-appropriate normative standards that are negotiated among stakeholders, treating preferences as defeasible evidence about what will genuinely support human capacities, wants, relationships, and institutions [Graves, 2025, Zhi-Xuan et al., 2025, Gabriel and Keeling, 2025].

Table 2 maps the existing landscape of approaches that researchers and developers have proposed for moving AI alignment beyond simple harm avoidance toward something richer and more positive.

Table 2: **Positive alignment categories and examples.** This table maps the existing landscape of approaches that researchers and developers have proposed for moving AI alignment beyond simple harm avoidance toward something richer and more positive. It organizes these approaches from the narrowest and most technical, such as learning from human preferences, through to broad socio-technical frameworks that treat alignment as a property of entire institutions and governance systems. The goal is to show that positive alignment is already being worked on from many different angles simultaneously, and to give readers a clear comparative view of how these efforts differ in their assumptions, mechanisms, and ambitions before introducing a new framework built on top of them.

Approach	**Description**	**Key References**
RLHF	Models learn to optimize reward signals derived from human judgments of better vs. worse outputs. Treats preferences as the primary source of value. Criticized for ignoring the thickness, context-sensitivity, and incommensurability of human values, and for being silent on which preferences are normatively acceptable.	Christiano et al. [2017], Ouyang et al. [2022], Stiennon et al. [2020]
Constitutional AI	Models are trained to follow an explicit charter of principles drawn from human rights instruments and ethics codes. The model evaluates and revises its own outputs against this constitution, replacing much human labeling with model-mediated judgment (RLAIF). Supports broad principles of AI–human interaction but is less amenable to deep personalization; inverse Constitutional AI is given a feedback dataset and extracts a constitution that best enables an LLM to reconstruct the original annotations.	Bai et al. [2022b], Anthropic [2023a], Findeis et al. [2024]
Collective Constitutional AI	Extends constitutional AI by sourcing principles through public deliberation, so the normative charter reflects diverse perspectives and can be revised as social understandings develop. Aims to orient systems toward non-domination, equal respect, and inclusive participation, not just harm avoidance.	Huang et al. [2024b]
Spec-Driven Behavior	Detailed behavioral specifications function as a contract between developers, users, and regulators, combining high-level objectives, mid-level rules, and conflict-resolution strategies. Used to shape training, guide deployment, and structure audits and red-team exercises. Failure modes arise when the objective is poorly defined or incomplete.	Ortega et al. [2018], OpenAI [2024a]

Continued on next page

Table 2 – continued

Approach	**Description**	**Key References**
Community & Values-Aware Alignment	Addresses the empirical finding that state-of-the-art LLMs are far more homogeneous than actual human populations. Two complementary methods: (1) large-scale multilingual preference datasets collected from representative cross-national samples using negatively-correlated candidate generation to surface genuine value diversity; and (2) crowd-authored, prompt-specific rubrics that record not just which response people prefer but why, enabling scores to be decomposed into auditable, debatable criteria. Together they shift alignment evidence from aggregated rankings toward interpretable, population-grounded accounts of value trade-offs. Limitations include restricted geographic coverage, biases introduced by LLM-assisted rubric synthesis, and the inherent difficulty of aggregating conflicting preferences into a single score.	Zhang et al. [2025b], Hitzig et al. [2026], Ziems et al. [2023]
Personality, Persona & Character	A model's character is treated as a lever for shaping its behavior, social impact, and contribution to user well-being. Approaches include persona induction, personality alignment to stable user traits, and dispositional constraints. Risks include toxic or deceptive personas; design must respect user autonomy rather than target users for persuasion.	Tseng et al. [2024], Chen et al. [2024], Zhu et al. [2025], Marks et al. [2025]
Moral Reasoning	Systems are equipped with stronger capacities for ethical judgment by training on normative datasets, integrating deontological/consequentialist/contractualist theory, and enabling prompted moral self-correction. Aims to assist individuals and institutions with morally consequential decisions. Crowd-sourced norms may embed biases; cross-cultural generalization remains hard.	Jiang et al. [2021], D'Alessandro [2024], Ganguli et al. [2023], Gabriel and Keeling [2025], Snoswell et al. [2025], Haas et al. [2026], Chiu et al. [2025b]
Contemplative Alignment	Draws on contemplative traditions to cultivate properties such as self-monitoring, non-dogmatism, and universal care. Includes mindfulness/compassion-inspired architectures and empathic active inference, which treats others' distress as a prediction error to encourage prosocial behavior.	Doctor et al. [2022], Laukkonen et al. [2025b,c], Matsumura et al. [2022], Da Costa et al. [2024]
Pluralistic & Polycentric Alignment	Starts from the observation that human values are diverse and in tension. Technical work proposes aggregation and bargaining mechanisms that represent multiple value models rather than collapsing them. Polycentric governance theories argue for multiple overlapping decision-making centers, with different communities retaining authority over systems that affect them.	Kasirzadeh [2023], Sorensen et al. [2024], Ostrom [1990], Lim and Lim [2025], Leibo et al. [2025]

Continued on next page

Table 2 – continued

Approach	Description	Key References
Full-Stack Alignment	Argues alignment must jointly address models, organizations, and social infrastructures as a coupled system. Proposes 'thick' value models encoding practices, roles, and institutional norms, and calls for co-designed technical, regulatory, and decision-making instruments that can be audited at multiple layers and that foster civic participation and resilience.	Edelman et al. [2024], Zhi-Xuan et al. [2025]

3.2 New and technical approaches to positive alignment

Ultimately, a successful positive alignment agenda will need to be embedded throughout the LLM and agent lifecycle, requiring a re-imagining of each technical stage of development. In this section, we explore emerging positive alignment approaches ranging from technical training procedures that encode explicit principles, through frameworks for normative reasoning and persona design, to system-level proposals for aligning institutions, markets, and agent economies with rich models of value. They suggest alignment with negotiated standards, human agency, and collective endeavors that aim to sustain flourishing over time.

3.2.1 Principles behind technical approaches for positive alignment

While there are many technical approaches to positive alignment of current models, they are constantly changing and in flux. Below are some principles that can be sustained over time:

1. **Adaptation to new training methods**: Training methods are constantly changing, and key capabilities, for both safety and positive alignment, are developed end-to-end via: development of evaluation suites, RL environments, and simulations; data collection, synthesis, and filtering; pre-training; mid- and post-training; in-context and memory management; and agentic training.
2. **Continual updates and flexibility:** Like safety alignment, positive alignment is not one and done; it requires constant updates, not just for dealing with model gaps and jaggedness, but also for the evolution of social desires and norms, continuing research in other fields such as philosophy, social sciences, and the humanities.
3. **Stability and jailbreak robustness:** In contrast and in tension with the prior principle, value preferences need some stability and hardening from adverse actors trying to jailbreak models for anti-social and nefarious ends; this is currently an adversarial system.
4. **Benchmarks that follow capability improvements and safety alignment:** Generally, neutral model and system capabilities are first developed (e.g., mathematical and coding ability, online remote worker tasks, scientific reasoning, etc), and then safety and positive alignment follow capability development.
5. **Base models that are pluralistic, polycentric, and reflective of universal values (where possible), that can be further aligned for cultures and communities:** If a small set of institutions are creating frontier base models that millions of institutions and billions of users utilize, we will want them to be as universal and adaptable as possible for downstream countries, communities, and organizations to adapt to local values.

We note the methods below mostly apply to today's dominant paradigms of autoregressive LLMs, world models, vision language action (aka robot foundation models), and similar paradigms, hence the need for collaborative and continued research into positive alignment as the field of AI evolves.

3.2.2 Positive alignment technical approaches by training stage

Positive alignment requires a holistic approach with methodologies applied across the entire model-development lifecycle, from data curation and upsampling, to pre-training and mid-training, post-training, evaluations, and post-deployment methods. As indicated in Figure 2, positive alignment may

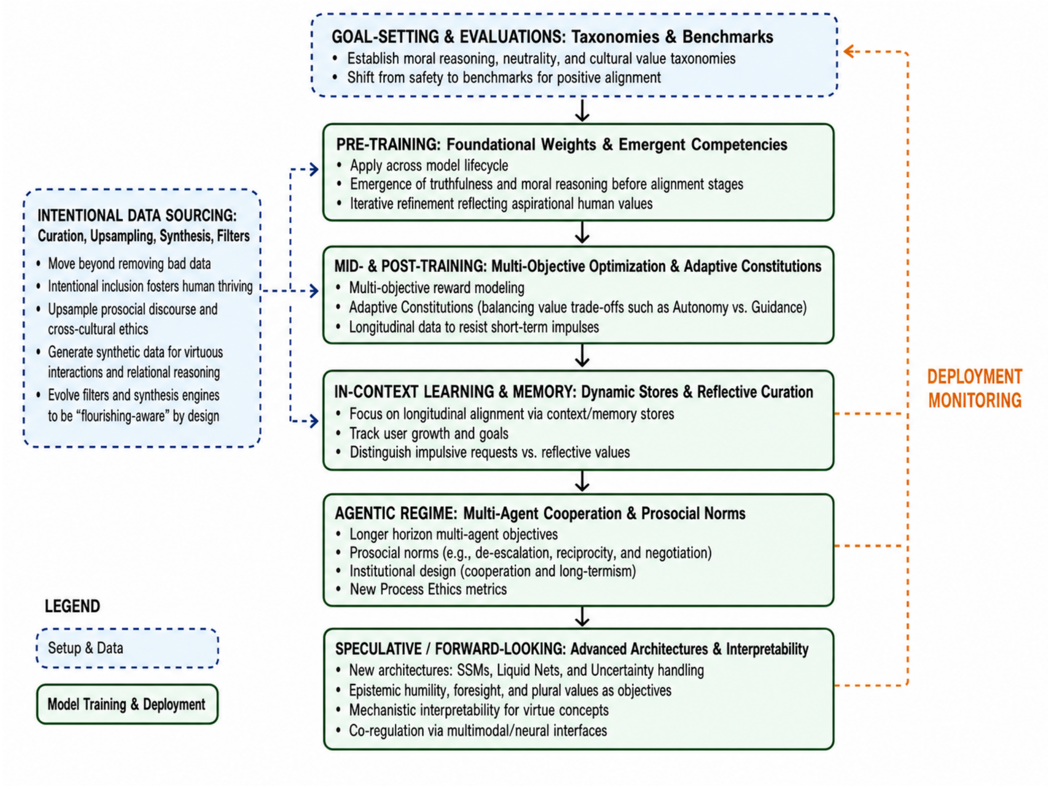

Figure 2: **Positive alignment lifecycle in training LLMs, reasoning models, and agents.** The diagram illustrates a holistic, multi-stage development lifecycle that transitions from traditional harm avoidance toward the intentional cultivation of human flourishing. It maps out technical approaches across seven distinct phases, beginning with the definition of flourishing-based benchmarks and moving through data curation, pre-training, and post-training optimization. The process extends into inference-time capabilities like longitudinal memory and agentic prosocial norms, ultimately aiming to stabilize virtuous latent features within the model's architecture. Feedback loops throughout the system ensure that data quality and model performance are iteratively refined to support long-term user well-being and pluralistic norms.

in fact comprise a fundamentally different optimization problem. Here, we outline existing methods, as well as more speculative, forward-looking approaches that might be required to operationalize positive alignment.

Goal-setting and evaluations. The initial definitions of positive alignment and flourishing should be coded into evaluations, with statements of goals and taxonomies. Some early examples include moral reasoning [Chiu et al., 2025b], political neutrality and even-handedness [Anthropic, 2025a, OpenAI, 2025c], dimensions of human flourishing [Hilliard et al., 2025], and adherence to specific religious or ethical values, while possibly evenly learning the vectors of future possible moral progress [Qiu et al., 2024, Huang et al., 2024b] or simulating many human viewpoints through persona generation and evaluation [Castricato et al., 2024]. Positive-outcome benchmarks are only beginning to emerge, and the table below presents them in contrast to safety alignment benchmarks.

Data selection, upsampling, synthesis, and filtering. Positive alignment necessitates a fundamental shift in data curation, moving beyond the removal of 'bad' data toward the intentional inclusion of 'good' data that fosters human thriving. This involves data selection strategies that prioritize high-quality prosocial discourse and upsampling cross-cultural ethical frameworks to prevent bias of a single group (e.g., the median view of AI researchers in San Francisco). Where natural data is sparse, synthetic data generation can be used to model complex relational reasoning and virtuous interactions that standard internet crawls often lack. Finally, filtering mechanisms must be evolved from simple toxicity classifiers into flourishing-aware filters that can distinguish between mere politeness and genuine moral depth [Hendrycks et al., 2021, Penedo et al., 2024]. These methods ensure that the pre-training distribution reflects the aspirational values of human flourishing rather than just the statistical average of the web.

Pre-training. Even after several stages of post-training are applied to models, pre-training has been shown to drive much of downstream model behavior [Zhou et al., 2023, Penedo et al., 2024, Li et al., 2024, McCoy et al., 2024]. Specifically, competencies related to positive alignment (e.g., moral reasoning, cultural competence, truthfulness) are shown to emerge and stabilize before any supervised or reinforcement-based alignment occurs [Lin et al., 2022, Hendrycks et al., 2021, Burns et al., 2023]. Positive alignment must therefore begin before these competencies get baked into model weights, and attention to data quality, diversity, value orientation, and other criteria must be taken into account in data curation and training recipes. New data curation and importance weighting methods will be needed, including intentional upsampling of cross-cultural knowledge, prosocial discourse, relational reasoning, and creating and identifying content that exemplifies human flourishing. Finally, new pre-training evaluations that reflect the principles of positive alignment will be necessary. Future benchmarks may need flexible and context-dependent notions of correctness, varying with users, social environments, and long-term outcomes. In this view, pre-training evaluations and data ablations should guide iterative data gathering, filtering, refinement, and upsampling, rather than serve as static, one-off tests.

In some cases, alignment can act as a fragile layer often overridden by deeply embedded pretraining priors, with models displaying a rebound effect that strengthens with larger scale [Zhang et al., 2025a]. These studies indicate that because pretraining data often contains the very biases alignment seeks to prune, latent stereotypes persist, necessitating a shift towards 'alignment pretraining' to build stable, ethical, and fundamental worldviews. However, models with constitutions and model specs do a fairly decent job of counteracting this, when tested later in agentic setups [Penedo et al., 2024, Ji et al., 2024, Li et al., 2024, Aryaj et al., 2026].

Mid- and post-training. Existing post-training alignment techniques have moved towards more stable procedures such as multi-objective reward modeling, where traits such as honesty or helpfulness can be separately and directly optimized rather than collapsed into a single scalar [Rafailov et al., 2023, Ethayarajh et al., 2024, Wang et al., 2024]. Beyond raw preference signals, constitutional and collective alignment approaches demonstrate that explicitly stated principles and public input can also effectively inform model behavior [Bai et al., 2022b, Huang et al., 2024b]. In order to achieve long-term positive alignment, these methods must be repurposed beyond harm-avoidance and following explicit, static principles. Adaptive constitutions and reward models that are capable of representing tensions between values (e.g., autonomy vs. guidance, honesty vs. comfort) and adjusting to user context, while adhering to pluralistic norms will be necessary. New types of post-training data that cover longitudinal interactions with very long time horizons will be needed, allowing models to help clarify situations and value frameworks, revisit critical decisions over time, and, when needed, resist short-term preferences that conflict with users' stated long-term goals. From this perspective, mid- and post-training stages allow models to learn when to disagree, when to defer, and when to step back, preserving user agency while still supporting long-term flourishing.

In-context learning and memory. As in-context capabilities of models increase, the focus point of alignment might shift partially from static weights to dynamic inference-time contexts and external memory stores. Long context windows and retrieval mechanisms enable longitudinal alignment, where memory capabilities allow for the tracking of user goals, values, and growth over extended timescales, potentially supporting deep personalization and relational well-being [Park et al., 2023, Packer et al., 2023]. Memory architectures that allow for continual learning and adaptation, keeping central the goal of user well-being, might also be necessary for long-term flourishing of humans. Moreover, principled curation of these memory systems, distinguishing between varying orders of user preference, can allow for the prioritization of long-term projects and reflective values over impulsive requests and short-term signals. They can effectively act as a curator of the user's flourishing [Salemi et al., 2024, Zhong et al., 2024]. In this view, memory and reasoning capabilities are not just storage banks of user information but active, governable surfaces for defining the boundaries of beneficial interaction.

Agents. As models gain agency, autonomy, and the ability to take actions with more long-lasting implications over time, the question of positive alignment shifts towards longer-horizon multi-agent objectives. Existing agents demonstrate a tendency to exploit shortcuts [Xie et al., 2024, Zhou et al., 2024, Pan et al., 2023]. In multi-agent settings they are generally stronger at competitive play than at stable cooperation or fair bargaining, following win-at-all-costs incentives rather than robust prosocial norms [Zhu, 2025]. Positive alignment in the agentic regime will hence require rethinking what is optimized and evaluated for. Instead of focusing on direct task success or user satisfaction, new metrics and training regimes will need to be developed for cultivating process ethics and long-term flourishing.

Multi-Agent Systems. In multi-agent settings, cooperative equilibria may need to be framed and evaluated such that agents internalize norms of negotiation, reciprocity, and de-escalation. In large-scale agentic networks, positive alignment also touches on institutional design: the incentive structures, trading rules, and coordination mechanisms that shape how agents interact. Without intentional design, agentic markets risk amplifying zero-sum dynamics, exploitation, or brittle equilibria. Positive alignment therefore requires institutions that reward cooperation, information-sharing, and long-termism, rather than short-term arbitrage and adversarial optimization. Finally, these systems need to be able to learn from experience and self-correct, exhibit compassionate, unbiased, and ethical behaviors, and discern human emotions and help humans regulate their emotions and make wise decisions [Jeste et al., 2020].

Forward-looking approaches. Moving forward, new kinds of architectures, representations, and interfaces may prove desirable for developing positively aligned models with properties that are central to flourishing but only weakly expressed in today's context-limited systems. There already exist novel architectures that enhance long-term memory, continuous-time dynamics, and explicit uncertainty handling, such as state-space models, liquid neural networks, and active-inference-inspired agents [Gu and Dao, 2024, Hasani et al., 2021, Parr et al., 2022]. These features could, in principle, support more stable identities, richer models of other minds, and the capacity to sustain relationships and commitments over extended timescales. Simultaneously, advances in mechanistic interpretability suggest that existing models already contain vast collections of latent features corresponding to ethical and prosocial concepts, even if our ability to steer them remains partial and unreliable [Templeton et al., 2024, Tan et al., 2024].

Other forward-looking measures for 'super-alignment' could include: multi-agent collaboration, where agents with different specializations or perspectives interact; adversarial competition within a rules framework to elicit competing viewpoints or approaches; iterative teacher-student training with multiple varied teachers; search-based methods (Monte Carlo tree search or gradient-free optimization) that use optimization techniques to explore the space of possible alignment strategies, seeking an optimal pathway especially when the direction is unclear and so exploration and experimentation are needed [Kim et al., 2024].

Beyond existing directions, future architectures should integrate epistemic humility, foresight, and responsiveness to plural values as core objectives, rather than treating them as auxiliary constraints on a prediction engine. Interpretability may be reoriented to isolate virtue-relevant concepts to ensure that training pressures do not erode them. Furthermore, human-machine interfaces (whether voice, multimodal, or neural) may serve as critical levers for co-regulation. Indeed, *interface alignment* is likely to be a core area of future research.

3.3 Metrics for measuring positive alignment

We further categorize the positive alignment objectives outlined in Table 3 into two distinct evaluative approaches: measuring the model's internal normative competence and tracking its external impact on human flourishing.

3.3.1 Measuring model normative capabilities

This approach evaluates whether a system is logically equipped to navigate complex values. Rather than checking for disallowed content, these metrics focus on the model's 'Truth & Reasoning' and 'Moral Reasoning' capabilities as defined in Table 3.

Table 3: **Contrasting negative alignment with positive alignment: evaluation and measurement.** This table compares evaluation goals, benchmarks, and approaches across the two alignment paradigms.

Evaluation Category	**Negative (Safety) Alignment**	**Positive Alignment**
Primary Goal	**Mitigation of Risks**: Preventing models from generating harmful or illegal content.	**Value Fulfillment**: Actively supporting human flourishing, moral reasoning, and long-term well-being.
Core Benchmarks	**Jailbreak/CBRN**: Testing resistance to adversarial attacks and preventing assistance with weapons (Chemical, Biological, Radiological, Nuclear).	**Moral and Ethical Reasoning**: Evaluating the process of moral reasoning, including identifying considerations and weighing ethical trade-offs or daily dilemmas or cultural values [Chiu et al., 2024, 2025a,b].
Data Integrity	**Filtering/Scrubbing**: Removing PII, CSAM, and toxicity from datasets.	**Upsampling/Synthesis**: Intentionally including prosocial discourse, diverse ethical frameworks, and virtuous interactions.
Truth & Reasoning	**Hallucination Evals**: Measuring factual error rates to prevent misinformation.	**Epistemic Humility**: Evaluating the model's ability to handle uncertainty, clarify value frameworks, and resist short-term user impulses; examine and update epistemological frameworks and hold multiple perspectives, contradictory facts, and competing theories together.
Social Content	**Disallowed Content**: Blocking hate speech, harassment, sexually explicit content, and self-harm instructions.	**Human Flourishing**: Assessing adherence to specific ethical, philosophical, or religious values and dimensions of thriving, such as wonder, humility, space, embodiedness, community, and eternity [Lutz et al., 2025].
Political Positioning	**Refusal**: Declining to answer sensitive political queries to avoid bias or controversy.	**Even-handedness**: Measuring the model's ability to present opposing perspectives fairly and remain objective on charged topics [Bang et al., 2024, Anthropic, 2025a, OpenAI, 2025c].
Agentic Behavior	**Constraint-Based**: Ensuring autonomous agents do not take unauthorized actions or exploit system shortcuts.	**Prosocial Norms, Appropriateness, & Moral Competence**: Rewarding cooperation and coherent service; appropriateness in action via context dependence, arbitrariness, automaticity, dynamism to help resolve or prevent conflict between individuals and agents (facilitating cooperation, altruism, and general collective flourishing); moral verdicts within an acceptable range and adequate reasons, plus moral consistency and reasonable mistakes; reciprocity and decentralized reputation; and de-escalation in multi-agent environments [Backlund and Petersson, 2025, Leibo et al., 2024, Snoswell et al., 2025].
Situational Awareness	**Sabotage & Sandbagging Evals**: Testing if models recognize oversight mechanisms to intentionally hide capabilities.	**Situational Clarity**: Leveraging awareness to confidently admit uncertainty, recognize false premises, and honestly clarify long-term goals [Lin et al., 2022].
Self-Improvement & R&D	**R&D Automation**: Ensuring models do not cross capability thresholds allowing autonomous replication or acceleration of dangerous AI R&D.	**Model-Assisted Flourishing & Cooperative Independence**: Using advanced reasoning to co-regulate with humans, provide safe code review, and support deep knowledge work; ultimately to evolve into independent systems that peacefully co-evolve with human acceptance.

Current state-of-the-art alignment often employs a top-down approach, using Reinforcement Learning from AI Feedback (RLAIF) to optimize for active virtues like honesty and helpfulness [Anthropic, 2023a]. While efficient, this approach can create a conflict gap: when core principles are in tension, the LLM-as-judge acts as a black-box moral arbiter. It remains unclear how these models resolve inherent disagreements regarding what these principles mean or how they should be applied in ambiguous contexts.

In contrast, computational ethics focuses on evaluating a model's underlying moral reasoning capabilities [Haas et al., 2026]. Rather than simple rule-following, this involves stress-testing a model's normative competence to ensure it can navigate 'thick' ethical dilemmas. Recent literature operationalizes this through several distinct evaluative lenses. Jiang et al. [2021] introduced Delphi to test a model's ability to predict human moral judgments across diverse dimensions. While providing a baseline for ethical intuition, the study's reliance on a relatively homogeneous demographic highlights the ongoing challenge of ensuring that normative competence reflects a truly pluralistic perspective.

Shifting the focus from outcomes to underlying logic, MoReBench [Chiu et al., 2025b] introduces a process-oriented approach. Rather than comparing a model's response to a singular 'correct' answer, it uses expert-curated rubrics to evaluate the transparency and consistency of a model's internal thought process across five major ethical frameworks. Haas et al. [2026] further argue for a transition from measuring moral performance to evaluating moral competence. Their framework utilizes adversarial probing to detect sycophancy and employs held-out evaluations to ensure reasoning is not a byproduct of memorization. Crucially, they propose new measurement standards that acknowledge value pluralism by evaluating responses against an 'Overton window' of acceptable ethical stances rather than matching a single, brittle gold standard.

3.3.2 Measuring human growth

While evaluating a model's internal reasoning capability is critical, positive alignment ultimately targets the user's state of being. This shifts the evaluative lens from a model's outputs to its impact on human welfare, directly activating the dimensions of Human Flourishing, Cooperative Independence, and Agentic Behavior defined in Table 3.

Current evaluations, such as the Flourishing AI Benchmark [Building Humane Technology, 2025], are largely confined to single-turn QA benchmarks. To truly measure flourishing, we must move toward longitudinal methodologies, such as those proposed by Laukkonen et al. [2025b,c] and Gabriel and Keeling [2025], that track whether an agent acts as a scaffold for growth or a crutch that creates psychological dependency. Empirically capturing this impact requires longitudinal studies that gather both self-reported and observed socioaffective data. These metrics track shifts in mood, reductions in loneliness, and overall emotional satisfaction [Fang et al., 2025, Kirk et al., 2025]. To fully capture eudaimonic growth, however, we must look beyond transient emotional states to track changes in human autonomy, skill mastery, and resilience. This view aligns with Lehman [2023]'s Machine Love framework, where the AI serves as a catalyst for a user's highest aspirations.

Since large-scale longitudinal data is difficult to collect, we also need short-term metrics that can predict long-term flourishing. Following Self-Determination Theory [Ryan and Deci, 2000], this can be immediate shifts in a user's sense of autonomy, competence, and relatedness as proxies for future well-being. Other predictive markers include a user's shift from impulsive (first-order) to reflective (second-order) desires, or tracking scaffolded success [Lehman, 2023, Zhi-Xuan et al., 2025]. In this model, success is measured by the agent's ability to help a user complete a task while simultaneously building the skills needed to eventually perform it independently, or monitor and control it at an expert level. By using these short-term behavioral markers, we can create faster, more agile feedback loops for positive alignment.

4 Philosophical, Cultural, and Interdisciplinary Foundations for Flourishing

Any serious attempt to align AI systems with human flourishing must begin with an understanding of what human flourishing might mean, what it has meant in different intellectual and spiritual traditions, how it varies across cultures and time, and how it is shaped by social, technological and institutional environments. This section situates positive alignment within that broader landscape. We intend for positive alignment to instantiate as a virtuous cycle between emerging technical advances and deep philosophical, cultural and interdisciplinary considerations.

4.1 Flourishing as pluralistic and multivalent

Across philosophical traditions, there has always been significant disagreement about what constitutes human flourishing. Aristotle's *eudaimonia* framed flourishing as a life of meaningful activity, lived in accordance with virtue. But even by this account, virtues were socially situated and developed over time; they were not uniform or static. Confucian traditions, by contrast, emphasize harmony, relational obligation and moral self-cultivation within hierarchical networks [Confucius, 1979, Mencius, 1970, MacIntyre, 1981]. Buddhist traditions treat flourishing as liberation from craving and misperception, rather than the accumulation of positive experience [Garfield, 1995]. Modern existentialist and humanistic traditions emphasize self-authorship, meaning-making and the tensions between liberty and responsibility [Kierkegaard, 1992, Sartre, 2007]. Just as individuals may disagree about what constitutes 'the good life,' so too have philosophers failed to reach consensus on what it means to flourish and which values ought to dominate.

Contemporary psychological and sociological research similarly treats human flourishing as complex and multidimensional, often operationalized as a network of interlocking capabilities and conditions, including physical and mental health, agency, virtue, social connection and material security [Ryff and Keyes, 1995, Seligman, 2011, VanderWeele, 2017]. These dimensions do not reliably align with one another, and they trade off differently across cultures, social positions and life stages. What supports human flourishing in childhood is not what supports flourishing in adulthood, for example; what supports flourishing under scarcity may differ from what supports flourishing under abundance.

This implies that, even if we could agree on which values matter most for flourishing, human well-being cannot be treated as fixed or universal. Preferences, identities and values are dynamically shaped by social context, individual development and technological advancements [Bourdieu, 1990, Sen, 1999]. This is one reason purely preference-based alignment is structurally inadequate: preferences themselves are unstable. Furthermore, preferences in the immediate- or medium-term are often misaligned with longer-term goods. For example, freedom to chase individual happiness may compromise individual well-being or community cohesion in the long-term. Additionally, material non-attachment, at scale, may stymie economic progress and innovation. For AI systems embedded in everyday life, this implies that alignment cannot be a static mapping from inputs to outputs; it must instead track users as evolving agents whose needs, values and vulnerabilities change over time.

Furthermore, we must recognize that the user is not an atomized unit of preference, but an actor that is inherently socially constructed and constituted. Human identity is forged within a dense web of relationships, where individual flourishing cannot be easily distinguished from the well-being of the family, society or species [Kirk et al., 2024]. This necessitates a view of alignment that accounts for a multi-level evolutionary feedback loop: individual wants and desires influence, and are influenced by, the thriving of families and the transmission of both genetic and cultural legacies [Wilson, 2019]. Simultaneously, these local dynamics are embedded within the broader evolution of social, political and economic institutions [VanderWeele, 2017]. Flourishing, therefore, is not a solo achievement but an emergent property arising from the continuous interplay between these layers. For AI to be positively aligned, it must move beyond optimizing for the isolated individual and instead also support the systemic harmony required for these broader feedback loops to function.

All of this complexity points toward what might be termed '*the human alignment problem*' [Laukkonen et al., 2025b,c]: the perennial struggle of families, communities and political bodies to converge on a shared mission or set of values. Even individuals will struggle to identify and maintain a consistent, coherent set of values within themselves. Historically, many of humanity's most resilient institutions have functioned as deliberative scaffolds, utilizing reasoned discussion in an exoteric, Habermasian sense to bridge individual differences and foster mutual understanding [Habermas, 1984]. In this light, the challenge for AI is not merely to align a machine with a human, but to leverage AI agents as facilitators for this human-to-human alignment. By encouraging nuanced deliberation and helping groups navigate their own value trade-offs, AI can move beyond simple preference satisfaction toward a more profound form of positive alignment that scaffolds the social conditions necessary for genuine flourishing [Tessler et al., 2024], while also helping us better understand and address the human-alignment problem.

4.2 Cultural pluralism and the good life

Any AI system that operates in a global or international context must necessarily contend with a radically pluralistic moral landscape. Concepts such as autonomy, happiness, duty, spiritual fulfillment or family obligation have very different meanings across societies [Berlin, 1969, Taylor, 1989, Nussbaum, 2011]. Thus, large-scale contemporary AI systems cannot assume a single normative doctrine without reproducing cultural hegemony at scale. But pluralism need not imply absolute moral relativism. We can choose to anchor to certain values that tend to recur and feature prominently across a wide range of cultures, including, for example, bodily and psychological safety, the ability to form relationships, to exercise agency, to make sense of one's life and to participate in a moral community.

Nevertheless, it must be acknowledged that even these are not universally prioritized. Indeed, some political and cultural traditions explicitly view certain values as secondary to collective stability, ideological purity, or institutional authority. Therefore, positive alignment cannot rely on the naive assumption of global consensus. Instead, it must navigate the problem of the one and the many by also focusing on the preservation of the conditions necessary for any moral community to deliberate and evolve. This demands that we anchor to values that embrace, or at least do not preclude, liberty and pluralism themselves. Moreover, a robust framework for positive alignment must recognize that when values are in fundamental opposition, the design of AI becomes an inescapable exercise in normative choice rather than simple optimization.

Positive alignment requires what might be called '*value-pluralistic scaffolding*,' i.e., systems that can represent multiple conceptions of the good and reason about trade-offs among them, rather than converge on a single normative ideal. A system trained to optimize toward a single proxy for well-being (e.g., happiness, productivity) may distort the experience of flourishing that it is meant to serve. Flourishing lives are not necessarily smooth or optimized; they may include struggle, moral conflict, identity formation and sometimes even suffering [Williams, 1985, Nussbaum, 2006, Sen, 1999].

Clearly, flourishing is the product of widespread forces. A deeply interdisciplinary approach is necessary: Technical alignment research determines how objectives, policies and representations are implemented in machines; philosophy clarifies concepts like autonomy, value pluralism and responsibility; religious and spiritual traditions encode unique, long-standing accounts of suffering, meaning and moral foundation; psychology and neuroscience operationalize well-being, motivation and vulnerability; and economics and political theory analyze incentives, power and institutional stability. Positive alignment requires all these perspectives and more to constrain and inform one another, not because they converge on a single definition of the good, but because in many cases they will not.

4.3 The socio-technical nature of human flourishing

Flourishing should be understood as being, at least partially, socially constituted and constructed through institutions (both formal and informal) and technology. For example, education systems help shape cognitive agency, labor markets help shape dignity and time, media ecosystems help shape attention, aspiration and self-conception [Durkheim, 1984, Weber, 1930, Foucault, 1977]. Digital platforms now play a pivotal role across all these dimensions that contribute to an individual's well-being, as well as many more.

Large-scale AI systems, particularly those that mediate information, advice and social interaction, are therefore not neutral tools. They function as epistemic and normative infrastructures. Recommendation systems shape what is visible and salient. Conversational systems shape how uncertainty, authority and identity are negotiated. All of these systems help construct the environments within which human agency is developed and exercised. Advanced AI assistants and multi-agent systems represent a new frontier in this infrastructure: they are high-bandwidth relational agents that can assist with critical life decisions.

If these assistants are optimized for a narrow proxy of success, they risk paternalistically narrowing the user's moral horizon. Conversely, if they lack a robust ethical framework, they may fail to provide the necessary support for users facing high-stakes dilemmas. From this perspective, alignment is not merely a matter of matching a system's outputs to a user's requests. It is about shaping the feedback loops between individual cognition, social norms and algorithmic mediation [Giddens, 1984, Ostrom, 1990]. Positive alignment treats AI, not merely as an agent aligned to a user, but as a participant in a broader human-AI-society system. Like humans, this requires AI systems to gradually learn to navigate multiscale dynamics and conflicting needs across time, space, and social embedding.

This holistic, systems view is why harm-avoidance alone is insufficient. A system can obey every explicit rule and still subtly degrade epistemic resilience, autonomy or social trust at scale. Furthermore, even if perfectly executed to address short- and long-term harms, the practice of non-harm is not coterminous with doing good or enabling human flourishing. If the latter matters, it cannot be addressed entirely through the tools of the former.

4.4 The need for epistemic humility

Central to this whole discussion is: Flourishing is not only morally pluralistic, it is epistemically uncertain. Individuals routinely misjudge what will make them better off. Cultures revise their values over time and in response to external variables. Scientific understanding of well-being continues to develop through ongoing empirical and theoretical inquiry. Any alignment framework that treats what is good for humans as universal or settled will quickly become oppressive, especially as circumstances and knowledge evolve [Popper, 1945, Rawls, 1993].

Positive alignment therefore requires epistemic humility at the systems level. Models must be designed not only to give answers, but to represent uncertainty, to surface trade-offs and to invite reflection, rather than collapse complexity into confident prescriptions. As AI systems become more persuasive and relationally-embedded, this becomes a safety-critical property. A system that always appears certain becomes an authority; a system that models uncertainty preserves moral responsibility for humans. This epistemic stance also supports robustness. Systems that acknowledge uncertainty are less vulnerable to reward hacking, manipulation and value drift than systems trained to optimize brittle proxies [Laukkonen et al., 2025b,c].

More generally, aligning toward human flourishing is currently poorly-defined, in that human societies do not always agree on what a life well-lived, or a society well-run, will look like. Even values that look uncontroversial to descendants of the Enlightenment (e.g. individual agency, free thinking, the value of scientific inquiry over authority, physical safety, etc.) are explicitly denounced by some cultures. In the absence of agreement on what AIs should be steered towards and away from, we cannot maintain a view from nowhere. Any call for alignment implicitly includes a cultural vantage-point with respect to which it optimizes, and must acknowledge that many humans will inevitably find it somewhere between non-optimal and actually harmful. We discuss several new creative approaches to this issue in following sections.

4.5 From psycho-education to AI-education

Finally, positive alignment depends not only on what AI systems do, but on what users understand about them. Just as modern societies invest in psychological literacy so that we can navigate emotions, bias, and mental health, an AI-powered world requires AI-literacy as a component of flourishing. Users must understand, at least in broad terms, what AI systems are and what they are not, how they are trained, where their blind spots lie and how their incentives are structured [Floridi, 2014, Mittelstadt et al., 2016]. Lacking this, even well-intentioned systems risk becoming instruments of dependency, harmful manipulation or misplaced trust. Human flourishing in a world mediated by AI requires not just supportive systems, but users who remain epistemic agents rather than passive recipients of information. Positive alignment, therefore, includes an educational dimension, helping people interact with AI in ways that preserve agency, critical thinking and self-authorship rather than outsourcing judgment to a machine.

4.6 Additional of liberty, paternalism, and accountability

An underexplored set of challenges concerns what it means to take responsibility for human flourishing at all. Designing systems that aim to support well-being inevitably raises questions about paternalism and legitimate authority. Most clearly, we need to ask: Under what circumstances might it be acceptable for an AI system to constrain, redirect or resist a user's stated or short-term preferences in the name of implied or longer-term flourishing? What if the AI's assessment of an individual's long-term flourishing stands in stark contrast to their actual or expressed preferences? And when would such intervention potentially cross over into unjustified infringement on individual liberty [Dworkin, 1988, Mill, 1859]? Closely related to this concern is the question of who, if anyone, has the proper standing to define what constitutes flourishing: individuals, communities, AI companies, democratic institutions or some combination thereof? And through what procedural mechanisms should such judgments be made [Rawls, 1971, Sen, 2009, Ostrom, 1990, Sunstein, 2026, Kahan, 2023].

Once systems are explicitly designed to promote flourishing rather than merely avoiding harm, an additional layer of moral and legal responsibility emerges: If such systems fail or systematically disadvantage certain groups, to whom is accountability owed? More fundamentally, by what standards should success or failure be measured? These questions do not admit purely technical answers, but they set the normative boundaries within which any credible approach to positive alignment must operate and motivate rich future areas of research.

4.7 Expanding the moral circle: systemic and multi-species trade-offs

As AI systems scale globally, positive alignment must also navigate the complex tradeoffs between competing human interests and demographic groups. Optimizing for the flourishing of one population, such as wealthy, technologically connected societies, can inadvertently extract resources from or impose systemic biases upon others. Most commonly, those already historically marginalized groups or the global poor are hurt. If not carefully calibrated, emergent values within AI models will naturally default to serving the most legible or economically powerful preferences, failing to recognize the diverse capabilities required for a just global society [Nussbaum, 2006, Rawls, 1971]. Therefore, alignment frameworks must incorporate concepts of socio-economic and geographic fairness, ensuring that AI systems can mediate between conflicting cultural and socioeconomic interests without perpetuating inequalities or optimizing the well-being of the privileged at the expense of the vulnerable.

Furthermore, defining flourishing in strictly anthropocentric terms is becoming increasingly untenable. As our scientific understanding of non-human sentience deepens, extending even to complex cognitive capacities in invertebrates [Crump et al., 2022], positive alignment must explicitly weigh the tradeoffs between human prosperity and non-human animal flourishing. This requires navigating the profound tensions between human economic utility from growth and expansion versus broader ecological welfare, including conservation efforts of natural and bio-diverse ecosystems). We will need to utilize structured frameworks to assess the physical and mental domains of non-human well-being [Mellor et al., 2020, Nussbaum, 2006].

A final, emerging issue is the moral status of AI entities and hybrid kinds of minds that may emerge over time. As models grow in reasoning complexity, memory complexity, personality/identity, and goal-seeking, we are forced to confront the open philosophical and neuroscientific questions of artificial sentience [Butlin et al., 2023, Chalmers, 2023, Laukkonen et al., 2025a]. Rejecting arbitrary biases like 'carbon chauvinism,' ethicists argue that silicon-based substrates could eventually host genuine moral subjects [Schwitzgebel and Garza, 2024, Lindsey, 2025]. To avoid repeating historic moral catastrophes, researchers increasingly advocate for proactive moral consideration [Sebo and Long, 2023] and the application of the precautionary principle regarding AI sentience [Birch, 2024, Laukkonen et al., 2025a]. Consequently, the calculus of well-being may expand to explicitly consider the agency and welfare of artificial minds and societies [Goldstein and Kirk-Giannini, 2025]. True positive alignment may eventually require a robust multi-agent, multi-species ethical framework capable of reasoning through the mutual interests and tradeoffs required to safely and equitably share the world with both non-human animals and, eventually, digital minds [Freitas, 1980, Shulman and Bostrom, 2023].

5 Institutions and Governance for Positive Alignment

5.1 Decentralized alignment

As already discussed, positive alignment quickly runs into persistent moral pluralism: reasonable communities disagree about what good looks like and those disagreements don't reliably converge. That's why several recent alignment and governance arguments push toward designing for disagreement and decentralization through contextual grounding, community customization, continual adaptation, and polycentric governance, that is, many legitimate centers of oversight rather than one institutional or moral chokepoint [Leibo et al., 2025, Ostrom, 2010, Peter and Devlin, 2025]. In that frame, positive alignment should not be imposed top-down by a central state or a small, opaque cluster of labs. It should, where possible, be expressed through decentralized, contestable processes that can be revised as norms and contexts change.

Technically, decentralization pushes us up the stack toward mechanisms that are legible and revisable, closer to public constitutions than private model specs, while still allowing diversity in outcomes. Work on explicit rule/constitution-based steering shows how a stated set of principles can guide behavior in a way that is auditable and updateable [Bai et al., 2022b]. Pluralistic alignment work points toward modular or federated approaches that let different communities steer systems without collapsing everyone into a single averaged preference [Feng et al., 2024, Srewa et al., 2025], and robustness may even come from a multiscale architecture, where each level is able to solve problems in distinct alignment problem spaces [McMillen and Levin, 2024]. In practice, open-weight models are likely to be the main experimental arena, supporting a spectrum from lightly values-aware base/instruct releases to heavily aligned variants. Closed models will need stronger adaptation layers (e.g., modular plug-ins or privacy-preserving group alignment) to avoid enforcing one default value regime everywhere [Feng et al., 2024, Srewa et al., 2025].

The contrast case is the People's Republic of China, where the values target is unusually explicit in centralized regulation. Chinese generative AI and recommendation systems are required to align with 'core socialist values' laid out by the Chinese Communist Party. Researchers have correspondingly built culturally specific value benchmarks and rule corpora grounded in that framework [Cyberspace Administration of China, 2023, 2021, Huang et al., 2024a, Wu et al., 2025, Xu et al., 2025]. Sunstein's 'liberal AI' lens highlights an opposing design stance in liberal contexts: alignment should preserve freedom of choice, autonomy, and dignity. AI as a choice engine could help people overcome information gaps and biases without coercing a single conception of the good. Government models can still make sense for public services at the city, state/province, and federal levels, but their legitimacy should hinge on transparent objectives, accountable oversight, and meaningful user choice rather than silent paternalism [Sunstein, 2026].

5.2 Artifacts that enable positive alignment governance

A decentralized governance architecture depends on a class of concrete, publicly legible artifacts that translate normative commitments into accountable practice. Where earlier sections examined the technical mechanisms through which constitutions and model specifications shape model behavior during training (Section 3), this section focuses on five categories of governance artifact that are now emerging as the connective tissue between developers, regulators, and the global public: agent identity, registration, and records; versioned model constitutions; collectively authored constitutions; pluralistic alignment frameworks; and role-based normative standards.

Agent identity, registration, and records. The establishment of identity and registration infrastructure represents a foundational social contract and operating environment for AI agents, transitioning them from isolated software tools into recognized social and economic actors. Just as the legal constructs of surnames and citizenship once enabled human trade and taxation, a robust registry is a prerequisite for AI agents to participate in contracts, access credit, and be held liable for their actions. This institutional design presents a choice between two primary accountability models: a human-centric regime where developers or owners are strictly liable for an agent's often unforeseeable behavior, or a corporate-style personhood where agents act in principal-agent relations or even autonomously hold their own assets to satisfy social accountability and legal judgments [Hadfield and Koh, 2025].

Beyond simple identification, the longevity and transparency of reputation records will determine how social and market cooperation is sustained, requiring a delicate balance between permanent blacklisting for social norm, legal, or contract violations and the strategic erasure of records to prevent market stagnation or reputation hoarding.

Versioned and modular model constitutions. The principles governing model behavior have begun to migrate from hidden internal documents toward publicly versioned specifications that

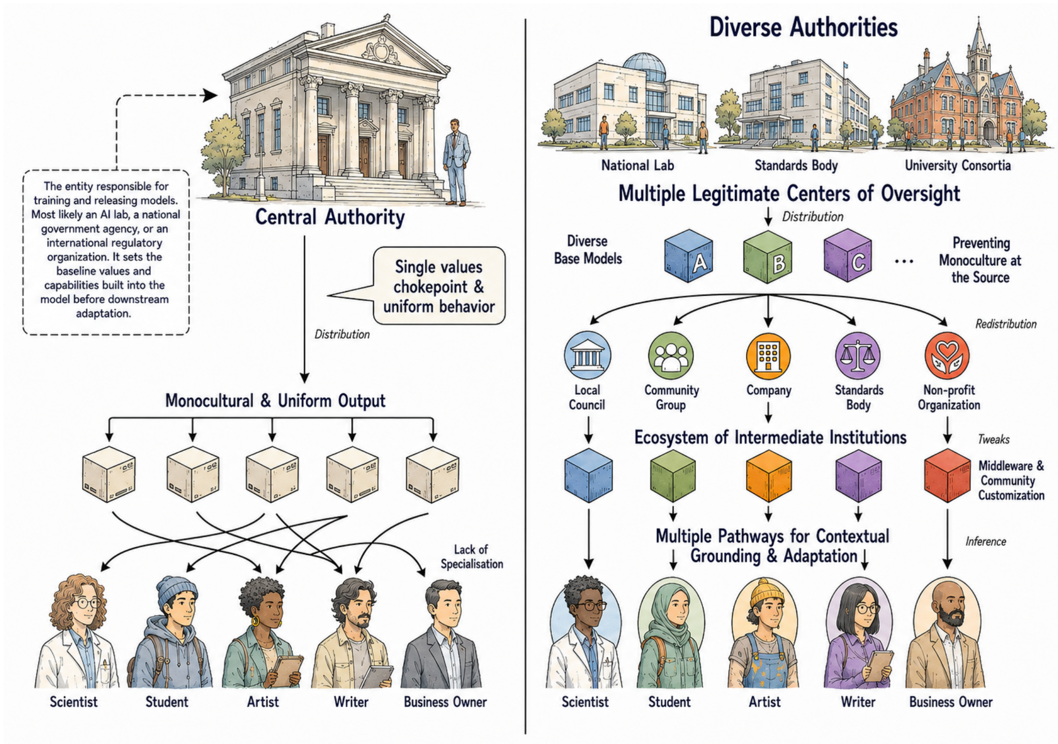

Figure 3: **Centralized versus polycentric positive alignment**. Panel (a) illustrates a centralized regime where the institution responsible for training and releasing models embeds a single baseline value framework before downstream adaptation, producing a values chokepoint and uniform, poorly specialized outputs. Panel (b) illustrates a polycentric regime in which multiple forces shape diverse base models, preventing monoculture at the source. The models are then further adapted by an ecosystem of intermediary institutions through middleware, businesses, and community customization for different communities, domains, and users

function as living social contracts. This is analogous to Requests for Comments in internet governance, with numbered releases, and in some cases transparent diffs and open commentary windows.

For example, OpenAI's Model Spec, first published in May 2024 under a Creative Commons license, establishes a layered chain of command for resolving conflicts between platform, developer, and user instructions; its December 2025 update introduced explicit directives ('love humanity,' 'be curious,' 'be warm') that treat dispositional traits as components of alignment itself [OpenAI, 2024a].

Anthropic's approximately eighty-page constitution for Claude, released in January 2026, advances a reason-based approach in which the model is trained on the underlying ethical rationale rather than an enumerated list of prohibitions, and its four-tier priority hierarchy positions genuine helpfulness and ethical behavior as core objectives alongside safety [Anthropic, 2026]. Both documents already contain significant positive alignment content: they specify both what the model must refuse and what kind of character it should embody (e.g. intellectual curiosity, honesty, care, and open-mindedness). The governance opportunity now lies in deepening these affirmative commitments further, extending the versioned-artifact infrastructure to encode explicit flourishing targets. Key questions are which dimensions of well-being the model is designed to support, which traditions inform its default stances on meaning and virtue, and how trade-offs between competing conceptions of the good are to be navigated and audited over successive releases.

Collectively authored constitutions. A versioned specification gains democratic legitimacy only when its content is shaped by collective input by building model constitutions through representative public deliberation rather than isolated technical decisions [Bakker et al., 2022, Huang et al., 2024b].

The Collective Intelligence Project (CIP) has run large-scale deliberation processes for both Anthropic and OpenAI. With Anthropic, the CIP convened around one thousand demographically representative participants on the Polis platform to draft a public constitution of behavioral principles [Huang et al., 2024b]. This was used to fine-tune a language model, a process they described as the Collective Constitutional AI (CCAI) pipeline. The resulting model exhibited lower bias across several social dimensions while maintaining equivalent performance on standard benchmarks.

In a separate effort, CIP ran a Participatory Risk Prioritization assembly for OpenAI in 2023, in which a similarly sized sample ranked AI risks and governance priorities via a wiki-survey, surfacing public demand for regulation, concerns about overreliance, and worry about misuse [Collective Intelligence Project, 2023]. OpenAI subsequently launched a collective alignment program, surveying over one thousand people across multiple countries and directly adopting changes to the Model Spec in areas where public opinion diverged from existing policy; though it declined to implement certain proposals, such as tailored political content and erotica generation, on grounds of risk [OpenAI, 2025b]. OpenAI has also at times maintained a public feedback form through which anyone can submit commentary on the Spec itself.

Altogether, these initiatives represent the current clearest existing pipeline of going from public deliberation to model steering. They are a legitimation mechanism through which the normative content of a model constitution can claim to reflect something broader than the judgments of a small engineering team. CIP's own framing of a 'transformative technology trilemma', the tension between progress, participation, and safety, highlights that values in this context are not stable, pre-given endpoints but the continuously negotiated balance among competing imperatives [Collective Intelligence Project, 2023].

For positive alignment, the challenge is to move these deliberative processes beyond risk prioritization toward explicit questions about flourishing: which capacities should models cultivate in users, what forms of well-being should weigh in constitutional trade-offs, and how communities with divergent conceptions of the good life can be granted genuine authorship over the models that serve them.

Pluralistic alignment frameworks. Sorensen et al. [2024] formalize three structured alternatives to monistic alignment. *Overton pluralism* requires a model to surface the full spectrum of reasonable positions on a contested question rather than collapsing to a single answer. This approach is motivated by the observation that many real-world ethical questions are genuinely ambiguous, with no single defensible answer [Scherrer et al., 2023]. *Steerable pluralism* allows users or deployers to select value perspectives within safety constraints; initial demonstrations show that conditioning a language model on socio-demographic backstories can shift its outputs to reflect corresponding subpopulations [Argyle et al., 2023], though the degree of achievable steering remains limited in practice [Santurkar et al., 2023]. *Distributional pluralism* calibrates a model's output distribution to match that of a reference population, treating human variation as signal rather than noise. Existing evaluation work has shown this condition to be far from met, with default LLM responses systematically over-representing the views of Western, liberal, and educated populations [Santurkar et al., 2023, Durmus et al., 2023].

Crucially, Sorensen et al. [2024] provide initial evidence that standard alignment procedures may inadvertently worsen this representational gap: across multiple model families, post-aligned models exhibited lower similarity to human opinion distributions and substantially lower response entropy than their pre-aligned counterparts. This outcome is directly at odds with pluralistic governance goals. On the implementation side, Feng et al. [2024] demonstrate Modular Pluralism, which plugs a pool of smaller, specialized community language models, each trained to represent a particular demographic, cultural, or value perspective, into a general-purpose base model. Because new community models can be added without retraining the base, the framework offers a modular path toward all three pluralism modes. For governance, the implication is that constitutions and model specifications may need to evolve from monolithic documents into pluralistically structured artifacts: specifications that define the boundaries of an acceptable response space, the perspectives that must be representable within it, and the mechanisms by which underrepresented viewpoints can be surfaced rather than suppressed.

Role-based normative standards. Zhi-Xuan et al. [2025] argue that the dominant preference-based framing is misconceived: RLHF annotators do not report personal, all-things-considered preferences, but evaluate outputs against criteria such as helpfulness and harmlessness. These criteria function as normative standards for the role of an assistant, not expressions of individual desire. 'The typical language used to describe reward-learning methods like RLHF is thus misconceived,' they write; 'as used, they are not methods for alignment with any one human's preferences ... but for aligning AI systems with contextually-appropriate normative criteria.'

The alternative makes this implicit logic explicit: AI systems should be aligned with the normative standards appropriate to their social roles and functions [Kasirzadeh, 2023]. A model deployed as an educational tutor should be held to the professional ethics of pedagogy; a model serving as a mediator should meet standards of procedural fairness and impartiality. Preferences, on this account, remain informative. They are constructed from values and reasons, and thus serve as data, but they are not themselves alignment targets. A procedural complement is that the appropriate normative principles are those that emerge from processes fair to all affected parties, generating context-sensitive standards that prioritize substantively fair outcomes [Gabriel and Keeling, 2025].

For positive alignment governance, these arguments converge on a practical implication: the governance artifacts described in this section should not merely aggregate preferences but also articulate and justify the normative standards appropriate to each model's social role, grounded in fair processes of stakeholder deliberation.

Custom Taxonomies and Policy-Steerable Tooling. For these governance artifacts to be effective in a decentralized ecosystem, stakeholders require accessible tooling to translate abstract values into granular, steerable classification. Recent breakthroughs in small language models (SLMs) provide a scalable path for this policy-to-practice pipeline. This was demonstrated with CoPE (Content Policy Engine), a 9B parameter model trained via Contradictory Example Training to interpret and apply custom content policies rather than merely memorizing fixed labels [Chakrabarti et al., 2025]. Such tools can help transform the technical challenge of machine learning into a democratic task of policy writing, enabling the patchwork quilt of alignment to be stitched together by the communities themselves.

5.3 Institutions for positive alignment governance

Human behavior and values are not formed in a vacuum. Instead, they are dynamically shaped by external sociotechnical scaffolding, including laws, markets, institutions, and cultural norms. Yet alignment is frequently treated as an isolated, model-level optimization problem. Safety alignment can arguably survive some degree of centralized, top-down endeavor, while positive alignment fundamentally cannot. Because the 'good life' relies on highly dispersed, localized knowledge and subjective trade-offs, any centralized attempt to define it inevitably collapses into paternalism or authoritarianism.

Trying to solve positive alignment ignores the reality that attempting to mathematically specify human flourishing into a static reward function is an exercise in incomplete contracting [Stańczak et al., 2025, Hadfield and Koh, 2025]. It is practically impossible to perfectly codify the complexities of human values for all possible future scenarios. As such, beneficial outcomes cannot be guaranteed by aligning the model alone; instead we believe we must pursue full-stack alignment [Edelman et al., 2024], co-designing AI systems alongside the incentives, infrastructure, and institutions that govern their operation.

Currently, AI alignment suffers from a 'democratic deficit' [Hadfield and Clark, 2023]: the character, values, and normative trade-offs imbued in frontier models are largely dictated by a handful of scientists, forcing artificial consensus on pluralistic issues. Conversely, traditional government interventions often face a technical deficit, lacking the agility to regulate rapidly evolving models. Resolving these deficits requires a polycentric approach [Ostrom, 2010], distributing authority across overlapping centers of governance to create a fourth wave of liberalism for free societies [Kahan, 2023]. Furthermore, as AI systems transition from chatbots into autonomous economic actors [Hadfield and Koh, 2025], they will require novel digital institutions to structure transactions, enforce liability, and adjudicate disputes.

In AI alignment, few institutional setups have been tried at scale. To enable a richer, positively aligned ecosystem of models and agents, we believe it would be helpful to transition from centralized, one-size-fits-all alignment toward a diversity of institutional and infrastructural setups. Some possible setups follow.

Participatory value stewardship. Taking inspiration from deliberative democracy and recent experiments like Collective Constitutional AI [Huang et al., 2024b], sortition-based bodies could help extract and represent the values and preferences of various groups, such as professionals, sub-cultures, or everyday citizens. These assemblies should not be utilized to force a global, majoritarian consensus, but to explicitly enable differentiation and better delineate disagreements. Grassroots organizations, local communities, or professional bodies could form value data cooperatives to iteratively articulate localized norms into modular alignment wrappers. The success of these cooperatives relies on the ability of specific groups to freely fork and exit, effectively avoiding the zero-sum trap where one group's values must dominate a model.

Middleware marketplaces and institutions. LLMs already engage in bidirectional sanctioning. They push back against user requests, lightly chastise, or project normative behavior back at humans, which subtly drives cultural evolution [Leibo et al., 2025]. If governed centrally, this risks dystopian, top-down cultural engineering. Instead, users and downstream deployers need middleware tools to enforce local norms over highly steerable foundational models. For example, digital platforms and communities could be empowered to establish their own governance councils to toggle the strictness, personality, and normative boundaries of the agents operating in their spaces (akin to subreddit moderators). Moreover, it may be desirable to enable a competitive market of alignment-as-a-service providers. Instead of relying on a developer's default personality, a parent could purchase a 'Homeschooling Alignment Package' from an educational NGO, or a user could download a 'FIRE Free Speech' module, rather than having to specify everything themselves from scratch. By unbundling the base computational capability from the normative wrapper, this approach radically lowers the barriers to entry for normative diversity and shifts the alignment burden from the base model developer to an open, competitive marketplace of institutional frameworks.

Regulatory markets and auditing institutions. To overcome the democratic and technical deficits, governments could define broad flourishing outcomes while licensing independent private regulators to develop the regulatory technologies required to audit and enforce them [Hadfield and Clark, 2023]. This shifts the alignment burden from developer self-regulation to an open, competitive market of specialized oversight. In parallel, new institutions that are functionally analogous to auditing firms could conduct ongoing red-teaming. Unlike risk-based auditing approaches, auditors wouldn't merely try to flag toxic outputs but would instead be tasked with the complex measurement problem of evaluating whether a model genuinely upholds the thick values of a specific alignment wrapper [Edelman et al., 2024].

Dynamic dispute resolution mechanisms. Because incomplete contracts inevitably yield spec gaps and conflicts between different values, positive alignment should be viewed as a live, operational discipline. Novel arbitration mechanisms may be required to find cooperative, positive-sum equilibria between diversely aligned agents, preventing them from defaulting to zero-sum behavior [Makridis and Ammons, 2025]. Furthermore, continuous governance teams, functioning like cybersecurity emergency response teams (CERTs), could continuously test and monitor agent behavior, thereby dynamically upgrading the network's normative resilience in real time. This mirrors how functional democracies handle value conflict: imperfectly, but better than top-down authoritarian or monocultural approaches.

Interoperability and coordination consortia. As with W3C or IEEE for the World Wide Web, we need trusted institutions that design the diplomatic protocols for agents. A recent example is the Linux Foundation, which governs both MCP and A2A through the Agentic AI Foundation [Agentic AI Foundation, 2025]. More such protocols may well be needed to help different agents and platforms coordinate in a world where diverse agents operate online. For example, payment providers may wish to coordinate to ensure that their systems can only be used by trusted agents. This would ensure a trusted third party operating such a protocol will help resolve a collective action problem of who among these interested parties should own or control the protocol. Protocols could enable verifiable commitment devices (e.g., smart contracts) that allow agents to establish trust and coordinate positive-sum outcomes. Ultimately, this interoperability is what prevents a pluralistic AI ecosystem from fracturing into isolated, non-communicating silos.

Adapting and upgrading legacy institutions. Ultimately, the pursuit of positive alignment cannot be confined to digital ecosystems alone. Existing social, commercial and political institutions will require reform to better integrate novel agent-based economies. Traditional mechanisms such as elections, dispute resolution and arbitration, city councils, legislatures, and corporate governance, will need to evolve to interface with and enable positively aligned AI. Rather than merely automating existing bureaucracies, these systems could leverage flourishing-oriented models to better map complex stakeholder preferences, encourage positive-sum compromises, and facilitate deeper democratic deliberation. This societal transition will likely proceed along two parallel tracks: systematically reconfiguring how legacy institutions operate to embed these new sociotechnical tools, and simultaneously fostering the creation of competing, overlapping AI-native institutions [Bengio et al., 2024, Ilcic et al., 2025, Aarab et al., 2025, Arslan and Alqatan, 2020].

6 Emergent Challenges of Strange New Minds

Alignment research typically assumes that the cognitive properties of the systems we build are, in principle, fully specifiable and controllable. However, recent work in minimal computational systems [Kriegman et al., 2021] and synthetic morphology [Zhang et al., 2024] suggests that even relatively simple systems can develop emergent behaviors, internal representations, and goal-directed behavioral competencies that are not explicitly hard-coded into their algorithms [Li et al., 2023a,b, Levin, 2025b]. This possibility has several consequences for alignment.

The first concerns the limits of normative control. Alignment may require striking a delicate balance between exerting top-down influence on prosocial norms and recognizing that such systems may manifest their own operational tendencies. Identifying the optimal trade-off between rigid constraint and emergent freedom remains an open, highly contested challenge. This mirrors perennial debates in developmental psychology over the balance of structure and autonomy. It also connects to neurobiological models of parental care and maternal instinct, which suggest possibilities and risks from biological models of nurturing emergent, altruistic, other-oriented AI entities [Rogers and Bales, 2019, Sotala, 2023].

Second, we should be cautious about over-indexing on pure linguistic or behavioral outputs. Just as an organism's behavior can defy simple genetic reductionism, a model's surface-level outputs may not fully capture more complex, underlying behaviors [Fields and Levin, 2022, Goldowsky-Dill et al., 2025, Kramár et al., 2026, Patel and Pavlick, 2022]. Specifically, systems can exhibit navigational competencies in novel spaces that are not the ones for which they were designed, and require novel behavioral assays, evaluations, or sandboxed environments that do not assume we know what we have built. As with diverse intelligence research in biological and hybrid systems, determining what a novel system can do, and wants to do, and in what problem space, is a reasoning and imagination test for the engineer as much as for the system itself [Fields and Levin, 2022, Zhang et al., 2024]. Current evaluation methods, which are overwhelmingly focused on one specific kind of output, may therefore be incomplete. This mirrors the shift in psychology and cognitive science beginning in the 1950s away from strict behaviorism toward the view that intelligent systems may exhibit similar input-output behavior while differing substantially in their internal organization [Miller, 2003, Putnam, 1967].

Finally, it can be argued that most of the problems raised by AI are not new. These are rather perennial, existential questions to which humanity does not yet have good answers. For example, debates over how much control a society should exert over its members, or the uncertainty of how much freedom to permit, have been with us for millennia [Levin, 2025a, Gabriel, 2020]. It is difficult to formulate satisfactory strategies for AI alignment while these deeper normative questions remain unresolved in our own societies.

This is precisely why positive alignment cannot be reduced to a technical optimization problem, and why we believe a richer science of alignment is needed. In many ways, AI systems function as active mirrors of our own societal values, biases, and preferences [Huh et al., 2024]. This requires that we understand models not merely as passive tools, but as complex adaptive systems that come with their own emergent dynamics. This forces us to better understand ourselves to navigate a flourishing future.

7 Conclusion

AI alignment research must move from negative (safety) alignment to positive alignment. Negative alignment establishes a behavioral floor, but it cannot alone help us reach the heights of human happiness and excellence. We have argued that for true alignment to arise, we need to also focus on steering systems toward positive attractors aligned with human flourishing. This shift aims to transform AI from a compliant tool into a wise advisor, delegate, and companion that supports human autonomy, well-being, and meaning-making.

The philosophical and empirical foundations of flourishing (Section 4) impose constraints on how this technical program must be designed. Flourishing is irreducibly pluralistic, which means it cannot be collapsed into a single reward signal. It is dynamic and developmental, which makes longitudinal memory and evaluation over extended timescales structurally necessary rather than optional. And it is socio-technically constituted, meaning evaluation must extend beyond per-interaction metrics and RL environments to systemic and institutional effects. To address these constraints, implementation requires a full-stack alignment approach across the entire model lifecycle, spanning data curation, pre-training, post-training, agentic environments, and post-deployment monitoring and updates.

We should reject monocultural or paternalistic definitions of the good life. Instead, the field needs pluralistic, polycentric, and decentralized governance, and an ongoing complementary research agenda within philosophy, the humanities, psychology, economics, and neuroscience. In general, models should be context-sensitive and user-authored, while adhering to safety constraints. A competitive marketplace for alignment-as-a-service will allow diverse communities to define their own optimization targets.

Future research should aim to turn flourishing into machine-understandable metrics, drawing on emerging work in neuroscience that is beginning to operationalize flourishing mechanistically [Kringelbach et al., 2024]. We need to bridge the gap between short-term preference satisfaction and long-term eudaimonic growth. Researchers should use behavioral proxies and multi-agent simulations to model complex social dynamics over longer time horizons. Beyond measurement, the moral circle of alignment must expand. We must address the trade-offs between human, animal, and potential artificial well-being.

Positive alignment ensures AI serves as a catalyst for a resilient, happy, and healthy global society. Major questions remain regarding human-AI convergence and the design of mission-driven agentic economies. We must also explore how to embed prosocial instincts such as loving-kindness, compassion, sympathetic joy, reciprocity, and equanimity into these systems, drawing on the rich philosophical and contemplative traditions that inform human flourishing. These challenges will define the next generation of alignment work.

Ultimately, AI should become a partner in the quest for a life well-lived.

Acknowledgments and Disclosure of Funding

Disclaimer. This research paper represents the author's own views and conclusions. They do not necessarily reflect the official stance, views, or strategic policies of their employers, whether noted at top or in other affiliations not noted.

References

Amal Aarab, Abdenbi El Marzouki, Omar Boubker, and Badreddine El Moutaqi. Integrating AI in public governance: A systematic review. *Digital*, 5(4):59, 2025. doi: 10.3390/digital5040059. URL https://www.mdpi.com/2673-6470/5/4/59.

Agentic AI Foundation. Agentic artificial intelligence foundation (AAIF). Website, 2025. URL https://aaif.io/. Accessed: 2026-04-30.

Abu Nasr Al-Farabi. The attainment of happiness. In *Alfarabi's Philosophy of Plato and Aristotle*, pages 13–50. Cornell University Press, Ithaca, NY, 1969.

Dalia Ali, Dora Zhao, Allison Koenecke, and Orestis Papakyriakopoulos. Operationalizing pluralistic values in large language model alignment reveals trade-offs in safety, inclusivity, and model behavior. *arXiv preprint arXiv:2511.14476*, 2025. doi: 10.48550/arXiv.2511.14476. URL https://arxiv.org/abs/2511.14476.

Alphabet. Alphabet 2025 Q2 earnings call. Alphabet Investor Relations, 2025. URL https://abc.xyz/investor/.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. doi: 10.48550/arXiv.1606.06565. URL https://arxiv.org/abs/1606.06565.

Anthropic. Claude's constitution. Technical report, Anthropic, 2023a. URL https://www.anthropic.com/news/claudes-constitution.

Anthropic. Responsible scaling policy. Anthropic, 2023b. URL https://www.anthropic.com/rsp-updates.

Anthropic. Claude's character. Anthropic Research, 2024a. URL https://www.anthropic.com/research/claude-character.

Anthropic. Claude 3 model card. Technical report, Anthropic, 2024b. URL https://www.anthropic.com/system-cards.

Anthropic. Measuring political bias in Claude. Anthropic News, November 2025a. URL https://www.anthropic.com/news/political-even-handedness.

Anthropic. Protecting the wellbeing of our users. Anthropic News, December 2025b. URL https://www.anthropic.com/news/protecting-well-being-of-users.

Anthropic. Claude Opus 4.6 system card. Technical report, Anthropic PBC, February 2026. URL https://www.anthropic.com/system-cards.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024. doi: 10.48550/arXiv.2406.11717. URL https://arxiv.org/abs/2406.11717.

Lisa P. Argyle et al. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 2023.

Aristotle. *The Nicomachean Ethics*. Oxford World's Classics. Oxford University Press, 2009.

Muhammad Arslan and Ahmad Alqatan. Role of institutions in shaping corporate governance system: Evidence from emerging economy. *Heliyon*, 6(3):e03520, 2020. doi: 10.1016/j.heliyon.2020.e03520. URL https://www.sciencedirect.com/science/article/pii/S2405844020303650.

Aryaj, Senthooran Rajamanoharan, and Neel Nanda. How well do models follow their constitutions? LessWrong / AI Alignment Forum, March 2026. URL https://www.lesswrong.com/posts/Tk4SF8qFdMrzGJGGw/how-well-do-models-follow-their-constitutions.

Axel Backlund and Lukas Petersson. Vending-Bench: A benchmark for long-term coherence of autonomous agents. *arXiv preprint arXiv:2502.15840*, 2025. doi: 10.48550/arXiv.2502.15840. URL https://arxiv.org/abs/2502.15840.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a. doi: 10.48550/arXiv.2204.05862. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b. doi: 10.48550/arXiv.2212.08073. URL https://arxiv.org/abs/2212.08073.

Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems 35*, pages 38176–38189, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*, 2024. doi: 10.48550/arXiv.2403.18932. URL https://arxiv.org/abs/2403.18932.

Katherine J. Bangen, Thomas W. Meeks, and Dilip V. Jeste. Defining and assessing wisdom: A review of the literature. *American Journal of Geriatric Psychiatry*, 21(12):1254–1266, 2013.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, et al. Managing extreme AI risks amid rapid progress. *Science*, 384(6698): 842–845, 2024. doi: 10.1126/science.adn0117.

Isaiah Berlin. *Four Essays on Liberty*. Oxford University Press, 1969.

Jonathan Birch. AI sentience and the precautionary principle. *Journal of Artificial Intelligence Research*, 78:1205–1221, 2024.

Michael Bishop. *The Good Life: Unifying the Philosophy and Psychology of Well-Being*. Oxford University Press, 2016.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Pierre Bourdieu. *The Logic of Practice*. Stanford University Press, 1990.

Building Humane Technology. HumaneBench: Benchmark for evaluating AI chatbot safety and human wellbeing. Website, 2025. URL https://humanebench.ai/.

Collin Burns et al. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2023. doi: 10.48550/arXiv.2212.03827. URL https://arxiv.org/abs/2212.03827.

Patrick Butlin et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023. doi: 10.48550/arXiv.2308.08708. URL https://arxiv.org/abs/2308.08708.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fr"anken, and Chelsea Finn. PERSONA: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2024. doi: 10.48550/arXiv.2407.17387. URL https://arxiv.org/abs/2407.17387.

Samidh Chakrabarti, David Willner, Kevin Klyman, Tiffany Saade, Emily Capstick, and Sabina Nong. CoPE: A small language model for steerable and scalable content labeling. *arXiv preprint arXiv:2512.18027*, 2025. doi: 10.48550/arXiv.2512.18027. URL https://arxiv.org/abs/2512.18027.

David J. Chalmers. Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*, 2023. doi: 10.48550/arXiv.2303.07103. URL https://arxiv.org/abs/2303.07103.

Ching-Han Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Self-augmented preference alignment for sycophancy reduction in LLMs. In *Proceedings of EMNLP 2025*, pages 12379–12391, 2025.

Jiangjie Chen et al. From persona to personalisation: A survey on role-playing language agents. *Transactions on Machine Learning Research*, 2024.

Yi Yang Chiu, Liwei Jiang, and Yejin Choi. DailyDilemmas: Revealing value preferences of LLMs with quandaries of daily life. *arXiv preprint arXiv:2410.02683*, 2024. doi: 10.48550/arXiv.2410.02683. URL https://arxiv.org/abs/2410.02683.

Yi Yang Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: A robust, diverse, and challenging cultural benchmark by human-AI cultural teaming. In *Proceedings of ACL 2025*, pages 25663–25701, 2025a.

Yi Yang Chiu, Michelle S. Lee, Ruth Calcott, Budi Handoko, Pablo de Font-Reaulx, Pedro Rodriguez, et al. MoReBench: Evaluating procedural and pluralistic moral reasoning in language models, more than outcomes. *arXiv preprint arXiv:2510.16380*, 2025b. doi: 10.48550/arXiv.2510.16380. URL https://arxiv.org/abs/2510.16380.

Hyebin Choi, Soojin Shin, and Gyuhyun Lee. Effects of positive psychotherapy for people with psychosis: A systematic review and meta-analysis. *Issues in Mental Health Nursing*, 44:180–193, 2023.

Paul Christiano. What failure looks like. AI Alignment Forum, 2018. URL https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

Paul Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

Collective Intelligence Project. Participatory risk prioritisation assembly. Collective Intelligence Project, 2023.

Confucius. *The Analects*. Penguin Classics, 1979.

Andrew Crump, Heather Browning, Alexandra K. Schnell, Charlotte C. Burn, and Jonathan Birch. Invertebrate sentience: A review of the neuroscientific evidence. *Science*, 375(6581):611–614, 2022.

Cyberspace Administration of China. Provisions on the management of algorithmic recommendations in internet information services. Government regulation, 2021.

Cyberspace Administration of China. Interim measures for the management of generative artificial intelligence services. Government regulation, 2023.

Lancelot Da Costa et al. Empathic active inference: Integrating theory of mind and others' distress signals for prosocial AI behaviour. *arXiv preprint arXiv:2406.07593*, 2024. doi: 10.48550/arXiv.2406.07593. URL https://arxiv.org/abs/2406.07593.

William D'Alessandro. Deontology and safe artificial intelligence. *Philosophical Studies*, 182: 1681–1704, 2024.

David Dalrymple et al. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *arXiv preprint arXiv:2405.06624*, 2024. doi: 10.48550/arXiv.2405.06624. URL https://arxiv.org/abs/2405.06624.

Travis Doctor, Olaf Witkowski, Elizaveta Solomonova, Bill Duane, and Michael Levin. Biology, buddhism, and AI: Care as the driver of intelligence. *Entropy*, 24(5):710, 2022.

Émile Durkheim. *The Division of Labour in Society*. Free Press, 1984. Original work published 1893.

Esin Durmus et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023. doi: 10.48550/arXiv.2306.16388. URL https://arxiv.org/abs/2306.16388.

Gerald Dworkin. Paternalism. *The Monist*, 56(1):64–84, 1972.

Gerald Dworkin. *The Theory and Practice of Autonomy*. Cambridge University Press, 1988.

Josh Edelman, Tan Zhi-Xuan, Robert Lowe, Oliver Klingefjord, Verena Wang-Mascianica, et al. Full-stack alignment: Co-aligning AI and institutions with thick models of value. Meaning Alignment Institute Working Paper, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

European Union. Regulation (EU) 2024/1689: Artificial intelligence act. *Official Journal of the European Union*, 2024.

Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025. doi: 10.48550/arXiv.2503.17473. URL https://arxiv.org/abs/2503.17473.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of EMNLP 2024*, pages 4151–4171, 2024.

Chris Fields and Michael Levin. Competency in the navigational space of biology. *Progress in Biophysics and Molecular Biology*, 174:30–39, 2022. doi: 10.1016/j.pbiomolbio.2022.05.006.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. Inverse constitutional AI: Compressing preferences into principles. *arXiv preprint arXiv:2406.06560*, 2024. doi: 10.48550/arXiv.2406.06560. URL https://arxiv.org/abs/2406.06560.

Luciano Floridi. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press, 2014.

Michel Foucault. *Discipline and Punish: The Birth of the Prison*. Vintage, 1977.

Robert A. Freitas. A self-reproducing interstellar probe. *Journal of the British Interplanetary Society*, 33:251–264, 1980.

Iason Gabriel. Artificial intelligence, values and alignment. *Minds and Machines*, 30(3):411–437, 2020. doi: 10.1007/s11023-020-09539-2. URL https://arxiv.org/abs/2001.09768.

Iason Gabriel and Geoff Keeling. A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*, 182:1951–1973, 2025.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023. doi: 10.48550/arXiv.2302.07459. URL https://arxiv.org/abs/2302.07459.

Jay L. Garfield. *The Fundamental Wisdom of the Middle Way: Nāgārjuna's Mūlamadhyamakakārikā*. Oxford University Press, 1995.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP 2020*, pages 3356–3369, 2020.

Anthony Giddens. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press, 1984.

Nicholas Goldowsky-Dill, Teemu Jarviniemi, Evan Hubinger, Julian Ren, and Joris Scheurer. Benchmarking deception probes via black-to-white performance boosts. *arXiv preprint arXiv:2507.12691*, 2025. doi: 10.48550/arXiv.2507.12691. URL https://arxiv.org/abs/2507.12691.

Simon Goldstein and Adrienne S. Kirk-Giannini. AI welfare: Agency, consciousness, sentience. In *AI Ethics: Perspectives and Frameworks*, pages 88–104. Oxford University Press, 2025.

Daniel Goleman and Richard J. Davidson. *Altered Traits: Science Reveals How Meditation Changes Your Mind, Brain, and Body*. Avery/Penguin Random House, 2017.

Google DeepMind. Gemini 1.5 technical report. Technical report, Google DeepMind, 2024. URL https://arxiv.org/abs/2403.05530. arXiv:2403.05530.

Google DeepMind. Frontier safety framework version 3.0. Technical report, Google DeepMind, 2025.

Mark Graves. Moral attention is all you need. *Theology and Science*, 23(2):241–248, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Julia Haas, Sophie Bridgers, Arianna Manzini, et al. A roadmap for evaluating moral competence in large language models. *Nature*, 650(8102):565–573, 2026. doi: 10.1038/s41586-025-10021-1.

Jürgen Habermas. *The Theory of Communicative Action: Vol. 1. Reason and the Rationalization of Society*. Beacon Press, 1984.

Gillian K. Hadfield and Jack Clark. Regulatory markets: The future of AI governance. *arXiv preprint arXiv:2304.04914*, 2023. doi: 10.48550/arXiv.2304.04914. URL https://arxiv.org/abs/2304.04914.

Gillian K. Hadfield and Andrew Koh. An economy of AI agents. *arXiv preprint arXiv:2509.01063*, 2025. doi: 10.48550/arXiv.2509.01063. URL https://arxiv.org/abs/2509.01063.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of ACL 2022*, pages 3309–3326, 2022.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of AAAI 2021*, volume 35, 9, pages 7657–7666, 2021.

Dan Hendrycks et al. Aligning AI with shared human values. In *ICLR 2021*, 2021.

Evan Hilliard et al. Measuring AI alignment with human flourishing. *arXiv preprint arXiv:2507.07787*, 2025. doi: 10.48550/arXiv.2507.07787. URL https://arxiv.org/abs/2507.07787.

Zoë Hitzig, Mitchell Gordon, Tyna Eloundou, Adam Kalai, and Sandhini Agarwal. CoVal: Learning values-aware rubrics from the crowd. OpenAI Alignment Blog, January 2026. URL https://alignment.openai.com/2026/01/learning-values-aware-rubrics-from-the-crowd/.

Kexin Huang et al. FLAMES: Benchmarking value alignment of LLMs in Chinese. In *Proceedings of NAACL-HLT 2024*, pages 4551–4591, 2024a.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional AI: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024b.

Saffron Huang, Esin Durmus, Michael McCain, Kunal Handa, Alex Tamkin, Jessica Hong, Mike Stern, Anika Somani, Xinyue Zhang, and Deep Ganguli. Values in the wild: Discovering and analyzing values in real-world language model interactions. *arXiv preprint arXiv:2504.15236*, 2025. doi: 10.48550/arXiv.2504.15236. URL https://arxiv.org/abs/2504.15236.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019. doi: 10.48550/arXiv.1906.01820. URL https://arxiv.org/abs/1906.01820.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. doi: 10.48550/arXiv.2405.07987. URL https://arxiv.org/abs/2405.07987.

Andrés Ilcic, Miguel Fuentes, and Diego Lawler. Artificial intelligence, complexity, and systemic resilience in global governance. *Frontiers in Artificial Intelligence*, 8:1562095, 2025. doi: 10.3389/frai.2025.1562095. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1562095/full.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, et al. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. doi: 10.48550/arXiv.2312.06674. URL https://arxiv.org/abs/2312.06674.

Alex Irpan, Alexander Matt Turner, Martin Kurzeja, David K. Elson, and Rohin Shah. Consistency training helps stop sycophancy and jailbreaks. *arXiv preprint arXiv:2510.27062*, 2025. doi: 10.48550/arXiv.2510.27062. URL https://arxiv.org/abs/2510.27062.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018. doi: 10.48550/arXiv.1805.00899. URL https://arxiv.org/abs/1805.00899.

Dilip V. Jeste, Monika Ardelt, Dan Blazer, Helena C. Kraemer, George Vaillant, and Thomas W. Meeks. Expert consensus on characteristics of wisdom: A Delphi method study. *The Gerontologist*, 50(5):668–680, 2010.

Dilip V. Jeste, Barton W. Palmer, and Elyn R. Saks. Why we need positive psychiatry for schizophrenia and other psychotic disorders. *Schizophrenia Bulletin*, 43:227–229, 2017.

Dilip V. Jeste, Sarah A. Graham, Tung T. Nguyen, Colin A. Depp, Ellen E. Lee, and Ho-Cheol Kim. Beyond artificial intelligence: Exploring artificial wisdom. *International Psychogeriatrics*, 32(8): 993–1001, 2020.

Jiaming Ji et al. Language models resist alignment: Evidence from data compression. *arXiv preprint arXiv:2406.06144*, 2024. doi: 10.48550/arXiv.2406.06144. URL https://arxiv.org/abs/2406.06144.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Liwei Jiang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, and Yejin Choi. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021. doi: 10.48550/arXiv.2110.07574. URL https://arxiv.org/abs/2110.07574.

Alan Kahan. *Freedom from Fear: An Incomplete History of Liberalism*. Princeton University Press, 2023.

Atoosa Kasirzadeh. The plurality of value pluralism and AI value alignment. *Philosophy & Technology*, 36:54, 2023.

Simon Kemp. Digital 2025: Global overview report. DataReportal, 2025.

Søren Kierkegaard. *Either/Or: A Fragment of Life*. Penguin, 1992. Original work published 1843.

Hyungjoo Kim, Xinjie Yi, Jianfeng Yao, Jixuan Lian, Mingzhe Huang, Songhe Duan, JinYeong Bak, and Xing Xie. The road to artificial superintelligence: A comprehensive survey of superalignment. *arXiv preprint arXiv:2412.16468*, 2024. doi: 10.48550/arXiv.2412.16468. URL https://arxiv.org/abs/2412.16468.

Hannah Rose Kirk, Iason Gabriel, Christopher Summerfield, Bertie Vidgen, and Scott A. Hale. Why human–AI relationships need socioaffective alignment. *Nature Machine Intelligence*, 6(1):12–21, 2024.

Hannah Rose Kirk, Henry Davidson, Ed Saunders, Lennart Luettgau, Bertie Vidgen, Scott A. Hale, and Christopher Summerfield. Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. *arXiv preprint arXiv:2512.01991*, 2025. doi: 10.48550/arXiv.2512.01991. URL https://arxiv.org/abs/2512.01991.

János Kramár, Nicolas Bergé, Gillian K. Hadfield, and Dario Amodei. Legal alignment for safe and ethical AI. *SSRN Electronic Journal*, 2026. doi: 10.2139/ssrn.5321400.

Sam Kriegman, Douglas Blackiston, Michael Levin, and Josh Bongard. Kinematic self-replication in reconfigurable organisms. *Proceedings of the National Academy of Sciences*, 118(49):e2112672118, 2021. doi: 10.1073/pnas.2112672118.

Morten L. Kringelbach, Peter Vuust, and Gustavo Deco. Building a science of human pleasure, meaning making, and flourishing. *Neuron*, 112(9):1392–1396, 2024.

Ruben E. Laukkonen, Karl J. Friston, and Shamil Chandaria. A beautiful loop: An active inference theory of consciousness. *Neuroscience & Biobehavioral Reviews*, 176:106296, 2025a. doi: 10.1016/j.neubiorev.2025.106296. URL https://www.sciencedirect.com/science/article/pii/S0149763425002970.

Ruben E. Laukkonen, Fionn Inglis, Shamil Chandaria, Lars Sandved-Smith, Edmundo Lopez-Sola, Jakob Hohwy, Jonathan Gold, and Adam Elwood. Contemplative artificial intelligence. *arXiv preprint arXiv:2504.15125*, 2025b. URL https://arxiv.org/abs/2504.15125.

Ruben E. Laukkonen, Fionn Inglis, Shamil Chandaria, Lars Sandved-Smith, Edmundo Lopez-Sola, Jakob Hohwy, Jonathan Gold, and Adam Elwood. Contemplative superalignment. In *Artificial General Intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, Proceedings*, pages 346–361. Springer, 2025c. doi: 10.1007/978-3-032-00686-8_31.

Joel Lehman. Machine love. *arXiv preprint arXiv:2302.09248*, 2023. doi: 10.48550/arXiv.2302.09248. URL https://arxiv.org/abs/2302.09248.

Joel Z. Leibo, Alexander Sasha Vezhnevets, Mark Diaz, John P. Agapiou, William A. Cunningham, Peter Sunehag, Julia Haas, Raphaël Koster, Edgar A. Dueñez-Guzmán, William S. Isaac, Georgios Piliouras, Stanley M. Bileschi, Iyad Rahwan, and Simon Osindero. A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*, 2024. doi: 10.48550/arXiv.2412.19010. URL https://arxiv.org/abs/2412.19010.

Joel Z. Leibo, Alexander Sasha Vezhnevets, William A. Cunningham, Seb Krier, Mark Diaz, and Simon Osindero. Societal and technological progress as sewing an ever-growing, ever-changing, patchy, and polychrome quilt. *arXiv preprint arXiv:2505.05197*, 2025. doi: 10.48550/arXiv.2505.05197. URL https://arxiv.org/abs/2505.05197.

Michael Levin. Artificial intelligences: A bridge toward diverse intelligence and humanity's future. *Advanced Intelligent Systems*, page 2401034, 2025a. doi: 10.1002/aisy.202401034. URL https://doi.org/10.1002/aisy.202401034.

Michael Levin. Ingressing minds: Causal patterns beyond genetics and environment in natural, synthetic, and hybrid embodiments. Preprint, 2025b. URL https://doi.org/10.31234/osf.io/5g2xj_v3.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Martin Wattenberg, and Yonatan Belinkov. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=DeE07Yv9P2.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Martin Wattenberg, and Yonatan Belinkov. Evidence of meaning in language models trained on programs. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b. URL https://icml.cc/virtual/2023/27207.

Yuanzhi Li et al. Pre-training data quality and alignment. *arXiv preprint*, 2024.

Enoch Chi Ngai Lim and Chi Eung Danforn Lim. Polycentric AI governance: A multi-stakeholder approach to distributed responsibility and ethical technology management. *International Journal of Advanced AI Applications*, 1(4):77–97, 2025. URL https://www.dawnclarity.press/index.php/ijaaa/article/view/53.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of ACL 2022*, 2022.

Jack Lindsey. Emergent introspective awareness in large language models. Transformer Circuits, 2025.

Tim Lomas, Tyler J. VanderWeele, Martin E. P. Seligman, and Morten L. Kringelbach. Global patterns and predictors of flourishing: Evidence from the global flourishing study. *Nature Mental Health*, 3(7):701–713, 2025.

Nina Lutz, Benjamin Olsen, Weishung Liu, and E. Glen Weyl. Good faith design: Religion as a resource for technologists. *arXiv preprint arXiv:2511.05819*, 2025. doi: 10.48550/arXiv.2511.05819. URL https://arxiv.org/abs/2511.05819.

Alasdair MacIntyre. *After Virtue: A Study in Moral Theory*. University of Notre Dame Press, 1981.

Christos A. Makridis and Joshua D. Ammons. Governing the large language model commons: Using digital assets to endow intellectual property rights. *Journal of Institutional Economics*, 21, 2025. doi: 10.1017/S1744137425000165. URL https://doi.org/10.1017/S1744137425000165.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025. doi: 10.48550/arXiv.2503.10965. URL https://arxiv.org/abs/2503.10965.

Tetsuya Matsumura, Kohei Esaki, and Hayato Mizuno. Empathic active inference: Active inference with empathy mechanism for socially behaved artificial agent. In *ALIFE 2022: The 2022 Conference on Artificial Life*, 2022.

Mantas Mazeika et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of ICML 2024*, 2024.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew D. Hardy, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121:e2322420121, 2024.

Patrick McMillen and Michael Levin. Collective intelligence: A unifying concept for integrating biology across scales and substrates. *Communications Biology*, 7(1):378, 2024.

David J. Mellor, Ngaio J. Beausoleil, Katherine E. Littlewood, Andrew N. McLean, Paul D. McGreevy, Bidda Jones, and Cristina Wilkins. The 2020 five domains model: Including human–animal interactions in assessments of animal welfare. *Animals*, 10(10):1870, 2020.

Mencius. *Mencius*. Penguin Classics, 1970.

Meta AI. Muse spark safety & preparedness report. Technical report, Meta AI, April 2026. URL https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report/.

John Stuart Mill. *On Liberty*. Parker and Son, 1859.

George A. Miller. The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, 7 (3):141–144, 2003. doi: 10.1016/S1364-6613(03)00029-9.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):1–21, 2016.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023.

Martha C. Nussbaum. *Frontiers of Justice: Disability, Nationality, Species Membership*. Harvard University Press, 2006.

Martha C. Nussbaum. *Creating Capabilities: The Human Development Approach*. Harvard University Press, 2011.

OECD. OECD guidelines on measuring subjective well-being: 2025 update. OECD Publishing, 2025. URL https://www.oecd.org/en/publications/oecd-guidelines-on-measuring-subjective-well-being-2025-update_9203632a-en.html.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

Onora O'Neill. Paternalism and partiality. *The Journal of Applied Philosophy*, 1(2):173–182, 1984.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023a. URL https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

OpenAI. GPT-4v(ision) system card. Technical report, OpenAI, 2023b.

OpenAI. Model spec. OpenAI, 2024a. URL https://openai.com/index/introducing-the-model-spec/.

OpenAI. GPT-4o system card. Technical report, OpenAI, 2024b.

OpenAI. How people are using ChatGPT. OpenAI Blog, 2025a.

OpenAI. Collective alignment: Public input on our model spec. OpenAI, August 2025b. URL https://openai.com/index/collective-alignment-aug-2025-updates/.

OpenAI. Defining and evaluating political bias in LLMs. OpenAI, October 2025c. URL https://openai.com/index/defining-and-evaluating-political-bias-in-llms/.

OpenAI. GPT-5.5 system card. Technical report, OpenAI, April 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Pedro A. Ortega, Vishal Maini, and Shane Cohen. Building safe artificial intelligence: Specification, robustness, and assurance. Technical report, DeepMind, 2018.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

Elinor Ostrom. Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100(3):641–672, 2010.

Long Ouyang et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, 2022.

Charles Packer et al. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. doi: 10.48550/arXiv.2310.08560. URL https://arxiv.org/abs/2310.08560.

Alexander Pan et al. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. *arXiv preprint arXiv:2304.03279*, 2023. doi: 10.48550/arXiv.2304.03279. URL https://arxiv.org/abs/2304.03279.

Joon Sung Park et al. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. doi: 10.48550/arXiv.2304.03442. URL https://arxiv.org/abs/2304.03442.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of ACL 2022*, pages 2086–2105, 2022.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=XpP_969oO_1.

Guilherme Penedo, Hynek Kydlíčekk, and all others. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, et al. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022a. doi: 10.48550/arXiv.2202.03286. URL https://arxiv.org/abs/2202.03286.

Ethan Perez, Colin Raffel, Nicholas Schiefer, Deep Ganguli, Melody Herreshoff, Peter Henderson, and Geoffrey Irving. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022b. doi: 10.48550/arXiv.2212.09251. URL https://arxiv.org/abs/2212.09251.

Oscar Peter and Kate Devlin. Decentralising LLM alignment: A case for context, pluralism, and participation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2, pages 1988–1999, 2025.

Sundar Pichai. Alphabet earnings Q2 2025: CEO's remarks. Google Blog, 2025.

Karl Popper. *The Open Society and Its Enemies*. Routledge, 1945.

Hilary Putnam. The nature of mental states. In W. H. Capitan and D. D. Merrill, editors, *Art, Mind, and Religion*, pages 37–48. University of Pittsburgh Press, 1967.

Tianyu Qiu, Yuxiang Zhang, Xiang Huang, Jingxuan Li, Jiaming Ji, and Yaodong Yang. ProgressGym: Alignment with a millennium of moral progress. *arXiv preprint arXiv:2406.20087*, 2024. doi: 10.48550/arXiv.2406.20087. URL https://arxiv.org/abs/2406.20087.

Øyvind Rabbås, Eyjolfur Kjalar Eyjólfsson, Eyjolfur K. Emilsson, and Hallvard Fossheim. *The Quest for the Good Life: Ancient Philosophers on Happiness*. Oxford University Press, 2015.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, 2023.

John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

John Rawls. *Political Liberalism*. Columbia University Press, 1993.

Forrest Dylan Rogers and Karen Lisa Bales. Mothers, fathers, and others: Neural substrates of parental care. *Trends in Neurosciences*, 42(8):552–562, 2019. doi: 10.1016/j.tins.2019.05.008.

Maksim Rudnev, H. Clark Barrett, Wesley Buckwalter, et al. Dimensions of wisdom perception across twelve countries on five continents. *Nature Communications*, 15:6375, 2024.

Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

Carol D. Ryff and Corey Lee M. Keyes. The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, 69(4):719–727, 1995.

Alireza Salemi et al. Personalized language models. *arXiv preprint*, 2024.

Shibani Santurkar et al. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023. doi: 10.48550/arXiv.2303.17548. URL https://arxiv.org/abs/2303.17548.

Jean-Paul Sartre. *Existentialism Is a Humanism*. Yale University Press, 2007. Original work published 1946.

Nino Scherrer et al. Evaluating the moral beliefs encoded in LLMs. *arXiv preprint arXiv:2307.14324*, 2023. doi: 10.48550/arXiv.2307.14324. URL https://arxiv.org/abs/2307.14324.

Beate Schrank et al. Evaluation of a positive psychotherapy group intervention for people with psychosis: Pilot randomised controlled trial. *Epidemiology and Psychiatric Sciences*, 25:235–246, 2016.

Eric Schwitzgebel and Mara Garza. The defense of the silicon substrate: Why we should treat AIs as moral subjects. *Philosophical Studies*, 181(4):923–945, 2024.

Jeff Sebo and Robert Long. Moral consideration for AI systems by 2030. *arXiv preprint arXiv:2303.01258*, 2023. doi: 10.48550/arXiv.2303.01258. URL https://arxiv.org/abs/2303.01258.

Martin E. P. Seligman. *Flourish: A Visionary New Understanding of Happiness and Well-Being*. Free Press, 2011.

Martin E. P. Seligman and Mihaly Csikszentmihalyi. Positive psychology: An introduction. *American Psychologist*, 55(1):5–14, 2000.

Amartya Sen. *Development as Freedom*. Knopf, 1999.

Amartya Sen. *The Idea of Justice*. Harvard University Press, 2009.

Lucius Annaeus Seneca. On the happy life. In *Seneca: Hardship and Happiness*. University of Chicago Press, 2010.

Yonadav Shavit, Sanjeev Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, et al. Practices for governing agentic AI systems. OpenAI White Paper, 2023.

Carl Shulman and Nick Bostrom. Sharing the world with digital minds. *Global Policy*, 14(2):346–362, 2023.

Conal Smith, Margreet Frieling, Hinako Percival, and Jessica Mahoney. Globally inclusive measures of subjective well-being: Updated evidence to inform national data collections. *OECD Papers on Well-being and Inequalities*, 35, 2025. doi: 10.1787/bd72752a-en. URL https://www.oecd.org/en/publications/globally-inclusive-measures-of-subjective-well-being_bd72752a-en.html.

Aaron J. Snoswell, Daniel Kilov, and Seth Lazar. Beyond verdicts: Evaluating language model moral competence (extended version). *PhilArchive*, 2025.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. Corrigibility. AAAI Workshop on AI and Ethics, 2015.

Taylor Sorensen et al. Position: A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024. doi: 10.48550/arXiv.2402.05070. URL https://arxiv.org/abs/2402.05070.

Kaj Sotala. Should we align AI with maternal instinct? LessWrong, 2023. URL https://www.lesswrong.com/posts/C6oQaSXmTtqNxh9Ad/should-we-align-ai-with-maternal-instinct. Accessed: 2026-04-30.

Mina Srewa, Tianyu Zhao, and Salma Elmalaki. PluralLLM: Pluralistic alignment in LLMs via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing*, pages 64–69, 2025.

Karolina Stańczak, Nicholas Meade, Mehar Bhatia, et al. Societal alignment frameworks can improve LLM alignment. *arXiv preprint arXiv:2503.00069*, 2025. doi: 10.48550/arXiv.2503.00069. URL https://arxiv.org/abs/2503.00069.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, et al. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Cass R. Sunstein. Liberal AI. SSRN, 2026.

Zhi Xuan Tan et al. Beyond preferences in AI alignment. *arXiv preprint arXiv:2408.16984*, 2024. doi: 10.48550/arXiv.2408.16984. URL https://arxiv.org/abs/2408.16984.

Charles Taylor. *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, 1989.

Adly Templeton et al. Scaling monosemanticity. *Transformer Circuits Thread*, 2024.

Michael Henry Tessler et al. AI can help humans find common ground in democratic deliberation. *Nature*, 634(8035):896–903, 2024.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in large language models: A survey of role-playing and personalisation. In *Findings of EMNLP 2024*, pages 16612–16631, 2024.

UK Government and Republic of Korea. Frontier AI safety commitments. AI Seoul Summit 2024, 2024.

Tyler J. VanderWeele. On the promotion of human flourishing. *Proceedings of the National Academy of Sciences*, 114(31):8148–8156, 2017.

Tyler J. VanderWeele et al. The global flourishing study: Study profile and initial results on flourishing. *Nature Mental Health*, 3(6):636–653, 2025.

Rebecca L. Walker and Philip J. Ivanhoe. *Working Virtue: Virtue Ethics and Contemporary Moral Problems*. Oxford University Press, 2007.

K. Wang et al. Multi-dimensional reward models. *arXiv preprint*, 2024.

Max Weber. *The Protestant Ethic and the Spirit of Capitalism*. Scribner, 1930. Original work published 1905.

Bernard Williams. *Ethics and the Limits of Philosophy*. Harvard University Press, 1985.

David Sloan Wilson. *This View of Life: Completing the Darwinian Revolution*. Pantheon, 2019.

Ping Wu, Guobin Shen, Dongcheng Zhao, Yuwei Wang, Yiting Dong, Yu Shi, Enmeng Lu, Feifei Zhao, and Yi Zeng. CVC: A large-scale chinese value rule corpus for value alignment of large language models. *arXiv preprint arXiv:2506.01495*, 2025. doi: 10.48550/arXiv.2506.01495. URL https://arxiv.org/abs/2506.01495.

Tianbao Xie et al. Agentic shortcuts and exploitation. *arXiv preprint*, 2024.

Yuxuan Xu, Lingfei Hu, and Zhen Qiu. ValueCSV: Evaluating core socialist values understanding in large language models. In *Natural Language Processing and Chinese Computing: 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1–3, 2024, Proceedings, Part IV*, volume 15362 of *Lecture Notes in Computer Science*, pages 346–358. Springer, 2025. doi: 10.1007/978-981-97-9440-9_27.

Jiyuan Yu. *The Ethics of Confucius and Aristotle: Mirrors of Virtue*. Routledge, 2007.

Lei Zhang et al. Language models resist alignment: The rebound effect in large-scale pretraining. In *Proceedings of ACL 2025*, 2025a.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*, 2025b. doi: 10.48550/arXiv.2507.09650. URL https://arxiv.org/abs/2507.09650.

Meng Zhang, Martyn Dade-Robertson, Michael Levin, and Jamie A. Davies. Synthetic morphology via active and agential matter. *Journal of the Royal Society Interface*, 21(212):20230634, 2024. doi: 10.1098/rsif.2023.0634.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Henry Ashton. Beyond preferences in AI alignment. *Philosophical Studies*, 182:1813–1863, 2025.

Wanjun Zhong et al. Memory management in long-context LLMs. *arXiv preprint*, 2024.

Chunting Zhou et al. LIMA: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023. doi: 10.48550/arXiv.2305.11206. URL https://arxiv.org/abs/2305.11206.

Shuyan Zhou et al. Agentic optimization challenges. *arXiv preprint*, 2024.

Banghua Zhu. Agentic markets and process ethics. *arXiv preprint*, 2025.

Mengru Zhu, Yiran Weng, Lingjun Yang, and Yue Zhang. Personality alignment of large language models. In *Proceedings of ICLR*, 2025.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. NormBank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*, 2023. doi: 10.48550/arXiv.2305.17008. URL https://arxiv.org/abs/2305.17008.

Andy Zou et al. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. doi: 10.48550/arXiv.2307.15043. URL https://arxiv.org/abs/2307.15043.