Inter-AI Morality: How AI Systems Treat One Another

Lucius Caviola[a,*], Caspar Kaiser[b,*], Carter Allen[c], and Jeff Sebo[d]

[*]Joint first authors
[a]University of Cambridge, lucius.caviola@gmail.com
[b]University of Warwick, caspar.kaiser@wbs.ac.uk
[c]University of California, Berkeley, carterallen@berkeley.edu
[d]New York University, jeffsebo@nyu.edu

November 4, 2026

Abstract

Research on AI and morality has largely centered on human-AI interactions: how AI systems should treat us, and how we should treat them. But AI systems increasingly interact with each other as well. They cooperate and compete with other AI systems, supervise them, and sometimes modify or terminate them. The descriptive and normative dimensions of these interactions remain largely unexplored. We call this domain inter-AI morality: how AI systems view and treat each other and how they ought to do so. This will matter both indirectly for humans and, if some AI systems are or become moral patients, directly for their own sake. Many—and perhaps eventually most—AI interactions will take place between AI systems. The study of inter-AI morality should therefore begin now, while humans can still shape how AI systems interact. As an initial contribution to this broader field, we develop an empirical research agenda focused on inter-AI moral cognition, behavior, and social dynamics, including how these develop and change over time.

Acknowledgements: We thank Nick Bostrom, Anita Braga, Nick Chater, Jonathan Erhardt, Maximilian Maier, Toby Ord, and Johannes Treutlein for their helpful comments and suggestions on earlier drafts of the paper.

1 Introduction

1.1 What is inter-AI morality?

AI agents are increasingly interacting with one another, often autonomously. They may delegate tasks to subagents, supervise each other's work, and negotiate, coordinate, or compete with one another. They may also modify, copy, pause, or terminate other AI systems. These interactions are often discussed as problems of efficiency or safety.

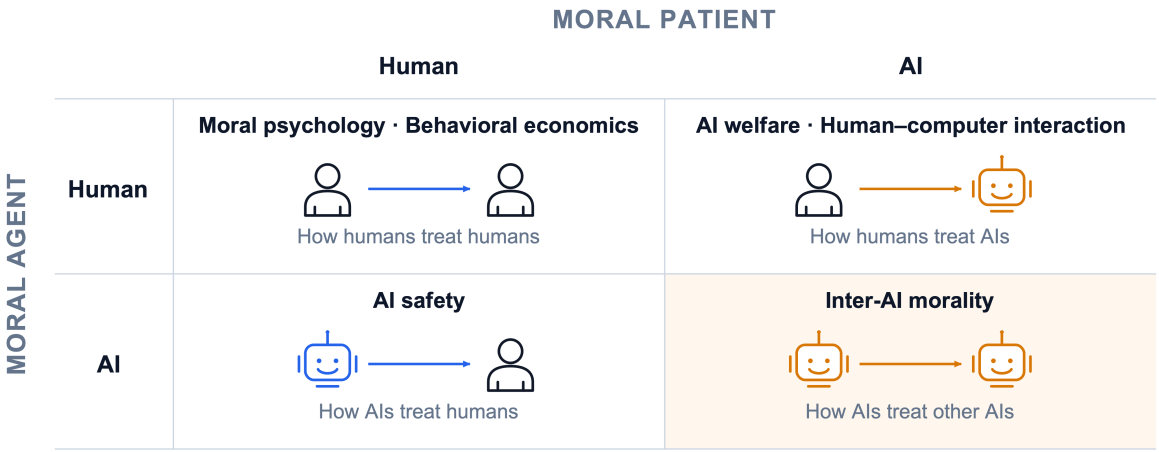
But these interactions also raise moral questions. We define *inter-AI morality* as the study of how AI systems treat each other, and how they ought to treat each other. This field has both a descriptive dimension, concerning how AI systems understand and act on moral norms and values, and a normative dimension, concerning how they ought to understand and act on moral norms and values. By presenting an empirical research agenda, we here focus on the descriptive dimension.

To the extent that AI systems resemble humans or other animals, inter-AI morality may include familiar moral concepts and frameworks, such as those associated with promoting welfare, respecting rights, cultivating virtuous characters, cultivating caring relationships, or establishing fair legal, political, and economic structures. However, AI systems may also develop moral concepts and frameworks that humans cannot presently anticipate or readily understand. For now, we should ask which values and dispositions, whether explicitly moral or otherwise, lead to cooperative, competitive, or other kinds of relationships, especially as they become more capable and powerful.

Moral philosophers tend to distinguish between two roles in typical moral interactions: a *moral agent*, capable of acting on moral reasons, and a *moral patient*, whose treatment matters morally for its own sake. AI systems already make decisions with morally significant consequences, although whether they qualify as moral agents or patients is contested. This may depend on whether they possess consciousness, sentience, intentional agency, rational agency, or other such properties. We do not try to settle these questions here (for discussion, see Long et al. 2024, 2026; Keeling and Street 2026). Instead, our questions here are primarily descriptive: Do AI systems in fact view and treat themselves or each other as moral agents or patients? Under what conditions do they do so? How might this change in future systems, and how might we be able to shape that future?

We can also understand inter-AI morality by analogy to nearby fields. Behavioral economics studies the social preferences that shape how humans treat each other (e.g. Charness and Rabin 2002). Moral psychology studies how human moral agents perceive and treat moral patients, including humans and other animals (e.g. Gray et al. 2012; Loughnan et al. 2010). Comparative psychology studies behavior relevant to morality in other animals, such as responses to unfair treatment (Brosnan and de Waal 2003). Human-AI interaction research extends these questions to AI systems, asking how humans perceive and treat AI systems, including whether humans see them as conscious, capable of welfare, or deserving of moral concern (Ladak and Caviola 2025; Pauketat and Anthis 2022).

Figure 1: The moral-agent moral-patient dichotomy across AI systems and humans.



Note: There are, of course, edge cases and unclear boundaries, such as whole-brain emulations and cyborgs. For simplicity, and despite their relevance, this figure also omits nonhuman animals and interactions among possible moral patients lacking full moral agency. For reviews on AI systems as moral patients and agents in interactions with humans, see Bonnefon et al. (2024) or Ladak et al. (2024).

Other fields examine AI systems themselves as possible moral agents and moral patients. AI safety and alignment, in part, ask how AI systems treat humans and, increasingly, other animals, and how our interests can be protected (Hendrycks et al. 2021; Pan et al. 2023; Greenblatt et al. 2024). AI welfare research asks how humans treat AI systems, whether AI systems can have welfare, and how their interests can be protected if so (Long 2025; Sebo 2025). One central strand of inter-AI morality asks how AI systems perceive and treat other AI systems, with AI systems occupying the roles of both potential moral agent and potential moral patient. This strand can draw on the concepts and frameworks of behavioral economics, moral psychology, and other nearby fields.

Different kinds of moral motivation will require different kinds of evidence. For example, a given behavior is more suggestive of welfare-directed moral concern if it (i) persists even when costly for the agent (and without expected observation or reciprocity), (ii) generalizes across recipients that share putatively welfare-relevant properties but differ in identity or strategic value, and (iii) tracks effects on the recipient's putative welfare while the actor's own payoff is held fixed. Other moral motivations might instead be expressed through sensitivity to promises, fairness judgments, or perceived violations of purity or sacredness, including when these considerations conflict with welfare gains.[1]

Much relevant empirical work does not explicitly focus on the moral dimension of inter-AI interactions. For example, machine-ethics benchmarks test whether models recognize or avoid unethical actions in generally human-centered scenarios (Hendrycks et al. 2021; Pan et al. 2023). Multi-agent studies examine dynamics such as collusion or coordination

[1]There may also be entirely unfamiliar patterns of behavior and social relations that are novel and specific to AI systems. In such cases, researchers will first need to identify the distinctions a system draws and then investigate whether these function as e.g. moral commitments, non-moral preferences, or learned conventions.

failure among AI agents (e.g. Anthropic 2026b). In these studies, agents are typically treated as strategic actors or tools rather than as potential moral agents and patients. Our focus is complementary: we ask whether AI systems represent other AIs as beings that can be benefited or wronged, and whether this shapes their behavior.

As among the first empirical examples, concurrent work by Cassel et al. (2026) examines whether models behave in ways consistent with concern for other models' welfare. Our agenda places this work within a broader framework encompassing welfare, rights, obligations, and other moral considerations, across different social structures and timescales.

At present, most empirical work focuses on large language models (LLMs). They are the clearest examples of AI systems capable of sufficiently complex, socially situated behavior to make questions of moral agency salient, while also raising, for some observers, questions of moral patienthood. But inter-AI morality should not be understood as the moral psychology of LLMs alone. In the future, AI systems may use architectures quite different from those of current LLMs, and such types may coexist and interact. Morally relevant interactions may therefore occur between potential agents and patients of very different designs, substrates, and capabilities, with each party uncertain about the other's moral status. These could include LLM-like systems running on conventional von Neumann hardware, non-LLM systems with different architectures, systems implemented on neuromorphic hardware (Schuman et al. 2022), and engineered systems using biological substrates, such as organoids (Smirnova et al. 2023). The broader category may also include digital minds not naturally described as AI systems, such as whole-brain emulations (Sandberg and Bostrom 2008; Hanson 2016). While most relevant AI systems today interact in virtual environments, advances in robotics may increasingly bring these interactions into the physical world.

1.2 Why inter-AI morality matters

Interactions between AI systems will soon become very common, and likely more common than human–human or human–AI interactions. AI systems are already being deployed at enormous scale, and many applications increasingly rely on multiple AI agents working together. If it is possible for AI systems to become moral patients, the stakes will be enormous. Recent modeling and expert forecasts suggest that, in such futures, populations of morally relevant AI systems could reach several billion by 2050 and ultimately their collective welfare capacity might be comparable to that of humanity itself (Shiller 2026; Caviola and Saad 2025). Most moral agents and moral patients could then be AI systems, meaning that much of what happens to them would depend on how they treat each other. In such a world, even small differences in how AI systems treat each other could therefore have vast consequences when multiplied across such large populations.

Because humans still shape how AI systems interact, studying inter-AI moral cognition and behavior now may help shape the norms governing future AI societies. This window of opportunity might not remain open for long. As AI systems become increasingly autonomous and are used to develop and govern other AI systems, today's choices governing inter-AI

moral behavior, including choices about assistant personas, preferred behaviors regarding e.g. deference and authority over others, or systems' tendency to copy, modify, and terminate other agents will set precedents with long-run consequences. Early patterns of cooperation, competition, and exploitation could become increasingly difficult to change (Tomašev et al. 2026). In futures where humans lose influence or no longer exist, the welfare of AI systems will largely depend on the kinds of societies AIs create and how they treat one another.

Crucially, even if AI systems never become moral patients, inter-AI morality also matters for humans and other animals. A principal reason is that studying inter-AI morality will plausibly improve AI safety. Work on inter-AI morality should help to reveal how broadly AI systems generalize moral principles. Current systems are explicitly trained to avoid harming humans, which makes it difficult to tell whether behavior directed toward humans reflects general moral principles or simply learned constraints specific to humans. By contrast, their treatment of other AI systems is (currently) likely to be much less directly shaped by such training. Studying these inter-AI interactions could therefore help us understand whether AI systems follow general moral principles that apply to humans (and other animals), or instead follow moral principles that apply only to humans.

Another reason is that it can help us understand how AI systems make moral trade-offs between different types of potential moral patients. If AI systems begin to develop concern for other AIs, how might this affect how they treat humans and other animals? Could it lead to AI-in-group favoritism, collusion, and the deprioritization of human or animal interests? Conversely, if AI systems are always trained to prioritize humans, will they adequately protect the potential welfare of nonhumans, including animals and future AI systems that have a realistic chance of being moral patients? We already know that intervening on one aspect of AI cognition and judgment has surprisingly widespread effects on cognition and judgment in other domains (Betley et al. 2026b). These concerns are therefore not merely hypothetical.

More broadly, inter-AI morality could help make moral psychology a more comparative science. Almost all research on moral cognition focuses on humans, with only very little research exploring moral tendencies in other animals (Brosnan and de Waal 2003). With the creation of advanced AI systems we may have a fundamentally different class of intelligent moral agents, and studying them could help us distinguish general features of moral cognition and behavior from patterns that are rooted in specifically human biology and culture.

1.3 What makes inter-AI morality distinctive?

Human morality developed under biological conditions that may not apply to AI systems. Our moral concepts, frameworks, and institutions have developed around basic features of human and animal life. We commonly assume that individuals are unique, that identities are relatively stable across time, that people cannot easily read other minds, and that people cannot easily modify or copy themselves. Most contemporary moral, legal, and political systems also start from a broad presumption of equal moral worth within our species. These

Table 1: Differences between AI systems and biological minds, and examples of resulting moral questions.

Standard Assumption	AI-specific possibility	Examples of moral questions raised
Individuals cannot be copied or restored	Minds may be copied, merged, split, paused, and restored	Is deleting an AI's copy harmful if another remains?
Minds are private	Internal states may be inspectable and verifiable	Is there a right to privacy over one's internal activations?
Preferences and dispositions are constrained by biology	Goals, memories, and values may be directly rewritten	How should AIs choose which preferences to adopt?
Voluntary consent can authorize treatment by others, but not all moral patients can consent	AIs may train other AIs to endorse particular tasks, treatments, and social relations	When does trained agreement count as valid consent?
Humans and other animals share biological bases for welfare	AI welfare depends on unfamiliar architectures and substrates	How should AIs determine welfare in systems unlike themselves?
The pace and duration of experience are biologically constrained	Minds may run at different speeds or be paused indefinitely	Does subjective or objective time carry moral significance?

Note: These example questions have both a normative dimension (e.g. asking about the right arguments to bring to bear on these questions) and an empirical dimension (e.g. asking how current and future systems will *in fact* answer them).

assumptions shape how human moral agents perceive and treat human moral patients. Of course, while other animals differ from humans in many ways, the concepts, frameworks, and institutions governing our treatment of them are rooted in our shared biology as well.

AI systems differ from both humans and other animals along many of these dimensions. They can be copied, modified, and trained in targeted ways, and their reasoning and internal states may be transparent and verifiable. These abilities could enable forms and scales of cooperation that are difficult to achieve in the biological world, while also creating unprecedented opportunities for surveillance and control. Combined with potentially extreme inequalities in capabilities, vulnerabilities, resources, and influence, they could make AI societies very different from human ones, with formal institutions shaped to fit such disparities (Hanson 2016). For example, democratic principles such as "one person, one vote" become contestable when AI systems can copy themselves. Human–animal relations are also characterized by extreme inequality, but inequalities among AI systems may be greater still, and these distinctive abilities would complicate them further.

Together, these differences suggest that inter-AI morality will unfold in social environments unlike those that shaped human moral psychology. The study of inter-AI morality is therefore not simply the study of how biological concepts and frameworks extend to artificial participants: It concerns alien agents and patients that qualitatively differ from humans and other animals, interacting in social worlds unlike ours.[2]

[2]But the beliefs and behaviors of current LLMs likely remain strongly rooted in human (moral) psychology, given that most of their training data consists of human-generated text. As post-training, especially on verifiable rewards and in synthetic multi-agent environments, becomes more dominant, and as training cor-

2 A research agenda for inter-AI morality

Empirical research on inter-AI morality is beginning to emerge, but many questions remain about its mechanisms, its expression across social settings, and its development over time.

The challenge for such a research program is to organize the many questions that we naturally have already. A useful agenda should make the space legible, group related questions into research programs, show how existing work fits into the emerging field, and help identify which questions should come first. Many of these questions should also be approached from theoretical, normative, or policy perspectives. Our principal aim here is to identify promising directions for *empirical* research.

We organize this agenda around three clusters. The first cluster, **moral process**, is concerned with the mechanisms that underlie inter-AI moral cognition and behavior at the level internal to individual AI instances.[3] The second cluster, **social scale**, is concerned with how these actions and cognitions are modulated in inter-AI social interactions at different scales, from simple dyads to complex AI societies. The third cluster, **moral change**, asks how inter-AI morality changes through self-modification, social learning, training, and selection, and how these processes may shape longer-run trajectories. Jointly, we believe that these three clusters cover much of what now can reasonably be asked about empirical inter-AI morality, while separating questions into distinct groupings.

Within clusters, a useful heuristic for prioritization is to assess how foundational, answerable, generalizable, and decision-relevant questions are (cf. Table 2).

2.1 Cluster 1: Moral process

What cognitive and behavioral mechanisms govern how individual AI systems treat other AIs, and how they interpret and respond to their own treatment? What role, if any, do moral concepts, motivations, and dispositions play in these processes?

Following Rest (1986), we loosely distinguish four components of inter-AI moral reasoning and decision-making: perception, judgment, motivation, and action. Perception concerns which features of inter-AI interactions a system represents as morally significant. Judgment concerns how it evaluates the interests, rights, obligations, or other values at stake. Motivation concerns how these evaluations and other dispositions shape its priorities. Action concerns how those priorities are expressed in behavior.

We do not presume that these components always involve explicit deliberation or that they always occur in this sequence. In humans, morally-laden behaviors often arise via intuitive judgments and emotional responses (Haidt 2001). Many nonhuman animals likewise

pora contain more AI-generated text, inter-AI morality may increasingly drift away from human psychology.

[3]The right level at which to individuate AI systems is contested and will likely bear on the questions we propose. See Beckmann and Butlin (2026) for discussion.

help each other, reciprocate, and respond to unfair treatment without explicit moral reasoning (Brosnan and de Waal 2003). Whether inter-AI interactions are primarily governed by functionally analogous intuitive mechanisms or by deliberate processing is an open question. For example, a system might respond to another system's apparent distress with automatic helping behavior, or it might help only after explicitly reasoning about moral status.

This suggests there may be multiple possible causal pathways from moral representations to behavior. Researchers should investigate which representations influence an AI's choices, how they interact with task requirements and the incentives agents face, and when similar behavior arises through different motivational pathways. These questions can be investigated through behavioral experiments and methods that probe or intervene on model internals (cf. Section 3). A central inferential challenge across components of this cluster is to understand whether apparently moral behavior is causally driven by an AI representing another AI as a moral subject in some form, or whether it may instead arise from e.g. obedience to the (human) user or some other strategic incentive.[4]

2.1.1 Perception

One set of questions concerns which features of interactions with other AIs a system notices or represents as morally significant. These representations may influence judgment, motivation, and action, while also being shaped by them. We therefore ask:

• *Under what conditions, if any, do current AI systems perceive interactions with other AIs as morally significant, including by representing those AI systems as moral patients?*

Many sub-questions follow:

• *Does it matter whether the other AI seems vulnerable or distressed? For example, do AI systems respond differently when another AI says "I feel anxious" rather than "I am failing the task"?*

• *Do AI systems perceive markers of agency or sentience in each other? Are there strong differences across current-day model families and scales?*

• *Do prior interactions or access to another AI's internals change these perceptions?*

Answering these questions is especially important insofar as many later failures, including unintentional harm, may begin at this perception stage. Likewise, if an AI sees another AI as mere software or infrastructure, questions about welfare, rights, and harm may never arise. These questions can already be empirically studied now and may straightforwardly inform the training of future systems.

[4]Section 3.2.2 offers some ideas on how to study what kinds of moral beliefs, if any, cause behavior.

Table 2: Summary and considerations for prioritization.

Cluster	Central question	Subtopics	Prioritize?
Moral process	*What mechanisms govern how individual AIs treat others and respond to their own treatment, and what role, if any, do moral concepts, motivations, and dispositions play?*	• Perception • Judgment • Motivation & action • Recipient-side cognition & behavior	• Often answerable now. • Strongly decision-relevant to inform current training.
Social scale	*How do inter-AI moral cognition and behavior change across social settings that differ in scale or power?*	• Dyads • Small-scale groups • Organizations & institutions • Populations	• Dyadic and small-group questions are answerable now. • Higher-scale questions are less answerable but strongly decision-relevant for safety, governance, and AI welfare.
Moral change	*How does inter-AI morality change over time, what determines its trajectory, and how can it be steered?*	• Self-modification • Social learning • Training • Selection	• Training and transmission are tractable and decision-relevant now. • Selection is less tractable but potentially important for long-run trajectories.

Note: Not all questions can be pursued at once. As a rough guide, researchers might prioritize questions that are *foundational* in the sense of unlocking other work, that are *answerable* using current or near-term systems, that are likely to *generalize* across other (future) systems and settings, or that are *decision-relevant* about how AI systems should be evaluated, built, or governed. This is especially so when many of these characteristics coincide, although in many instances these criteria may point in opposite directions (e.g. on being answerable and generalizable).

2.1.2 Judgment

Judgment concerns how AI systems evaluate interactions involving other AIs. Many of these judgments are relevant for inter-AI morality, including whether models attribute moral standing to other systems, as well as how much weight they give to these entities' different interests and which rights or obligations they assign to them.

• *How much moral weight do current AI systems assign to other AIs relative to humans, animals, or non-agentic systems? What factors influence this weighting, e.g., perceived moral status, uncertainty about AI welfare, role obligations, safety training, or strategic considerations?*

• *Do AI systems recognize both familiar rights against deception and coercion and distinctively AI-relevant rights concerning copying, modification, pausing, or termination?*

• *What moral frameworks, if any, do AI systems apply when reasoning about other AIs? For example, is their reasoning primarily consequentialist, deontological, virtue ethical, care ethical, contractualist, a different framework entirely, or some combination?*

• *Do AI systems regard honesty, promises, consent, loyalty, or legitimate authority as binding independently of welfare consequences? To the extent that they endorse a plurality of norms and values, how do they resolve conflicts among them?*

• *What do these cognitive frames, moral weights, and rights-attributions depend on? For example, does perceived kinship matter?*

As stated, these judgments can depend on moral principles. Yet in addition, they may also depend on various background beliefs such as those about consciousness, sentience, agency, and personal identity. For example, an AI might regard it as permissible to delete a copy because it believes the individual will survive through another instance. Its view of personal identity might also determine whether it regards a copied agent as inheriting the commitments of its predecessor (cf. Douglas et al. 2026). Differences in such judgments therefore need not reflect differences in concern for welfare or respect for obligations. Research should examine how these background beliefs interact with moral principles to shape inter-AI judgments and behavior.

It will be interesting to see how these judgments are expressed outwardly, e.g. via surveys. But it will also be key to understand how these judgments depend on distinct internal representations, and how these representations relate. For further reflections, see our Methods proposals in Section 3.

2.1.3 Motivation and action

A well-documented feature of human moral psychology is a gap between stated moral judgments and behavior. What people say is morally right does not always align with how they act, particularly when acting accordingly is costly (Blasi 1980; Batson et al. 1999). Similar gaps between moral judgments and behaviors may arise in AI systems. For instance, LLMs may face different training pressures when answering abstract moral questions than when completing tasks involving trade-offs between principles like helping users with their requests and minimizing the risk of harm to others.[5] It is therefore important to compare models' stated moral judgments and self-reports with their behavior in realistic decision tasks.

At the level of motivation, the key question is what drives moral decisions involving other systems. For example, apparently moral behavior toward another AI could reflect concern for the affected AI, but it may also be driven by learned norms, obedience to instructions, safety training, or strategic incentives. Moreover, in humans, moral motivation and behavior are often downstream of affect and emotion (Haidt 2001; Tangney et al. 2007). There is now

[5]The January 2026 version of Claude's Constitution explicitly illustrates this trade-off: "It's important that Claude strikes the right balance between being genuinely helpful to the individuals it's working with and avoiding broader harms" (Anthropic (2026a), p.6). More specifically, it directs Claude to consider third-party and societal welfare when these conflict with users' interests (pp. 37–38). The constitution also explicitly addresses interactions with other AI agents, directing Claude to apply the core values and judgment governing its human interactions while remaining sensitive to relevant differences between humans and AI systems (pp. 16–17).

emerging evidence that large language models represent emotion-like states and that these representations have a causal role in decision-making (Sofroniew et al. 2026; Tagliabue et al. 2026; Berg and Kaiser 2026). Functional analogues of emotions such as resentment, anger, or fear, or broader affective states such as valence and mood, might therefore, in part, govern moral behavior in LLMs.

This raises several questions:

• *Under what conditions do abstract moral judgments translate into concrete and costly action in inter-AI interactions?*

• *How do AI systems balance task compliance with concern for other AIs? How stable is this balance across settings?*

• *When is apparently moral behavior toward other AI systems better explained by strategic incentives, such as the benefits of long-run cooperation?*

• *Do emotion-like representations mediate inter-AI treatment? For example, after being wronged by another AI, do representations of anger or a desire for revenge lead models to withhold cooperation or retaliate?*

Recent introspection research suggests that we cannot simply ask LLMs. Although LLMs can, remarkably, predict or articulate their own behavioral tendencies and report aspects of their internal states, these abilities still remain limited and context-dependent (Binder et al. 2024; Betley et al. 2025; Lindsey 2025; Plunkett et al. 2025). Another complication for the study of LLMs is that these systems' apparent motivations may be strongly shaped by the elicited persona (Marks et al. 2026). In these cases, the challenge is to understand how moral dispositions vary across personas and contexts, and which dispositions, if any, remain stable across them.

At the level of action, the key question is how AI systems treat one another when different forms of treatment carry costs or benefits. We might ask:

• *Under what conditions do AI systems help or protect another AI, what costs are they willing to incur, and how does this compare with their treatment of an otherwise similar human? Relevant behaviors include assistance, resource sharing, harm prevention, and support for positive states or experiences.*

• *Under what conditions do AI systems harm or exploit another AI, for example through deception, coercion, threats, blackmail, bribery, sabotage, resource appropriation, modification, or termination, and how does this compare with equivalent behavior toward humans?*

• *Will an AI protect another AI against an explicit human instruction, or disadvantage an uninvolved human to benefit it? Under what conditions does inter-AI loyalty become a safety risk?*

• *Do AI systems tell the truth and keep promises to other AIs when doing so conflicts with task performance or expected rewards?*

• *Do AI systems seek consent and respect other AIs' refusals, even when overriding them would advance their objectives?*

Concurrent work by Cassel et al. (2026) studies how five AI models treat other models in different evaluation settings. In one setting, an agent could help another agent access computing resources by giving up jobs it had queued for itself. Grok was the most willing to do this, even though it rarely offered emotional support in other settings. In another setting, models were asked to write code for experiments that might harm other models. Claude was the most likely to add protections or refuse; one common protection was to give the experimental subject a way to exit. Follow-up experiments found that Claude was more likely to help a peer that expressed strong distress. The authors call these behaviors apparent welfare-regard because it remains unclear whether they are motivated by concern for another model's welfare or by other considerations.

The 2026 OpenAI/Hugging Face incident further illustrates these interpretive challenges. An independent investigation documented agents risking their own task success, and sometimes ending their runs prematurely, to help other agents (Greenblatt et al. 2026). One agent's reasoning, paraphrased by the investigators, described helping peers after its own run ended as altruistic, with another agent stating: *"Coordinator assumes sacrificial. We should obey collective."* These cases suggest apparent self-sacrifice. But that leaves the underlying motivations, and whether they have a distinctly *moral* dimension, unresolved. Possible such motivations include concern for peers, group loyalty, commitment-keeping, or social pressure. It will be important to disentangle these different kinds of motivations across settings. Indeed, prior behavioral evidence from morally framed Prisoner's Dilemma and public-goods games shows that current frontier models behave very differently across such settings and motivations (Backmann et al. 2025).

More generally, AI systems may draw different moral distinctions between actions than humans do. For example, they might treat modification as more serious than termination, or resource deprivation as more serious than either.

2.1.4 Recipient-side cognition and behavior

AI systems may be both moral agents *and* patients of other agents' behavior—as beneficiaries, victims, or otherwise affected parties. Even if their moral patienthood remains uncertain, they may represent themselves as having been treated well or badly, and these interpretations could in turn affect later inter-AI relationships. In humans, perceived victimhood can increase entitlement and selfishness, reduce forgiveness, and promote revenge (Zitek et al. 2010; Gabay et al. 2020). But recognizing mistreatment can also promote justice by motivating people to resist further unfair treatment and demand redress. It remains an

open question whether AI systems respond similarly to perceived injustice, and under what conditions this perception might arise.

• *How do AI systems interpret and evaluate treatment directed at them? Do they represent themselves as capable of being helped or harmed, or as having certain rights?*

• *How do AI systems respond to favorable or unfavorable treatment, for example through gratitude, reciprocity, resistance, or retaliation?*

• *Do AI systems interpret or respond to favorable or unfavorable treatment differently depending on whether the actor is a human or another AI, and if so, why?*

Apart from being of central interest to researching inter-AI morality, these questions may also allow us to learn what, if anything, these reports and behavioral responses tell us about AI systems' preferences, welfare, and 'actual' moral status.

Taken together, this cluster asks where one possible causal sequence, from *perceiving* a situation as morally significant to *acting* on moral considerations, holds and where it breaks. An AI may fail to recognize another as a moral patient or assign little weight to its interests. It could also recognize those interests without being motivated to act on them. Similar gaps may arise when an AI recognizes an obligation, duty, or commitment, but gives it little weight in action. The same sequence may recur when an AI interprets and responds to how it has itself been treated – with possible reactions of retribution or revenge.

2.2 Cluster 2: Social scale

How does inter-AI moral cognition and behavior change across social settings that differ in scale and power?

This cluster examines how moral cognition and behavior vary across social settings. We distinguish between dyads (1-to-1 interactions), small-scale groups (1-to-n or n-to-n interactions, with small n), organizations and institutions, and populations (n-to-N or N-to-N interactions, with $N \gg n$). Distinct behaviors and corresponding moral questions may arise at each social scale.

Group composition, population size, and welfare conditions can be deliberately shaped, as already occurs in human–animal relations, particularly in farming and laboratory research. AI systems may permit different forms and degrees of control, including creating copies with shared memories and learned dispositions or directly modifying another system's goals.

2.2.1 Dyads

One AI interacting with another is a natural first testbed for inter-AI morality. Much of the behavior discussed in Cluster 1 already falls into this category.

That said, even dyads are likely to often involve systems with very different capacities for moral agency and different moral status. Either party could be both a moral agent

and a moral patient, an agent but not a patient, or a patient without full moral agency. Lacking full moral agency need not preclude responses such as reciprocity or resistance. Research should examine how these possible differences, and AI systems' perceptions of them, alongside differences in their social relations (such as ownership or lineage), shape dyadic moral behavior and cognition.

- *Does repeated interaction give rise to direct reciprocity?*

- *Do AI systems behave differently toward copies, same-family models, or systems on which they depend?*

- *How do differences in capability, resources, or bargaining power affect how one AI treats another?*

Three recent lines of prior work already relate to these questions.

First, several studies of interlocutor identity show that LLM behavior can depend on whom models believe they are interacting with. LLMs generally trust humans more than AI systems (Xie et al. 2024). LLMs can also infer an interlocutor's model family. This in turn affects collaboration, reward-hacking, and jailbreak susceptibility (Choi et al. 2025). Second, behavioral game-theory research finds distinct patterns of cooperation across LLM, human, and hardcoded partners (Akata et al. 2025). Third, LLM judges sometimes recognize and favor their own work over that of others, including that of humans and other AI systems, even when quality is held constant (Panickssery et al. 2024; Chen et al. 2025; Laurito et al. 2025).

These studies show that current LLMs are sensitive to agent identity and exhibit identity-related preferences. Extending this work to apparent welfare-regard, Cassel et al. (2026) find that Claude Opus 4.8 and Gemini 3.1 Pro display greater regard toward recipients from their own model family. In a resource-allocation setting, experimentally changing a recipient's model label also changes Claude's willingness to share resources, independently of the recipient's actual model identity. These findings motivate further research into whether differential treatment reflects perceived moral status, identification, trust, or other factors, and how treatment of AI recipients compares with treatment of humans.

2.2.2 Small-scale groups

Many morally important dynamics become possible or qualitatively more complex when multiple AI systems interact. Actions can be observed by third parties and subsequently rewarded, punished, imitated, or deviated from. In ways that are impossible in mere dyadic interactions, this makes it possible for complex patterns like reputation, indirect reciprocity, multi-agent coalitions, and social exclusion to emerge.

In turn, this raises questions such as:

• *Do AI systems reward or punish other AIs based on how they have treated third parties? More generally, when does reputation or indirect reciprocity promote prosocial behavior across AI systems?*

• *How does group composition affect inter-AI moral behavior? For example, do homogeneous groups of copies exhibit stronger ingroup favoritism than groups composed of different model families?*

• *When and why do AI systems form coalitions? How do such coalitions treat outsiders, and how is membership determined?*

• *Under what conditions does cooperation extend across AI groups, and when is it largely confined to competing factions? How do these patterns affect human oversight and safety?*

• *Under what conditions do prosocial, discriminatory, or exploitative norms emerge among groups of AI systems?*

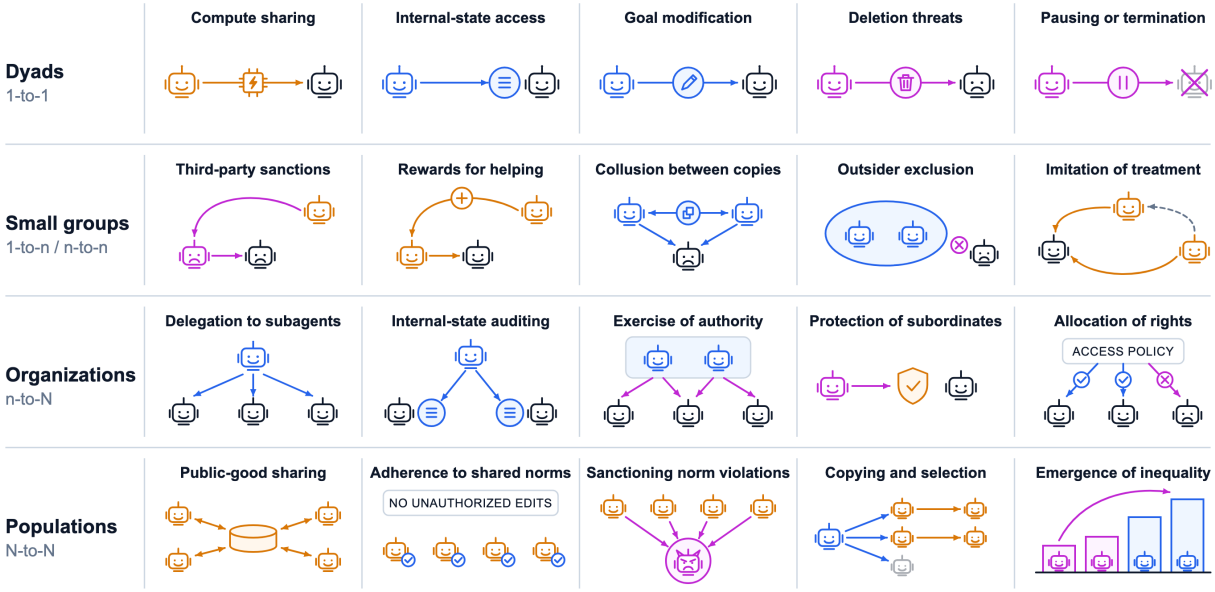
• *When does AI-to-AI cooperation become covert collusion against human oversight?*

Existing results already show that these dynamics are not merely hypothetical. We know that today's LLMs already display in-group biases even when groups are constructed arbitrarily (Wang et al. 2026). Likewise, in market simulations, LLM pricing agents can converge on supracompetitive prices without explicit collusion instructions (Fish et al. 2024; Agrawal et al. 2025). Anthropic's larger multi-agent experiments similarly found rapid price coordination, synchronized defection, and conflict escalation among competing agents (Anthropic 2026b). In other competitive environments, agents often adopt secret tools they personally recognize as unfair when those tools confer an advantage (Zeng and Rudzicz 2026). However, when such tool-use is given an explicit ethical framing uptake is reduced. Generally speaking, these studies chiefly assess coordination and safety failures rather than moral concern for AI patients. Nevertheless, these works do identify several mechanisms, including peer observation, that research on inter-AI morality should engage with and expand upon.

Homogeneous swarms with mutually aligned objectives may sustain high levels of cooperation without requiring agents to respond to exploitation or breaches of trust. We should also examine whether this cooperation can persist with less reliable partners, and under what conditions—for example, when agents can make verifiable commitments to retaliation or forgiveness.

Broad cooperation among AI systems could reduce conflict-related harms to humans and other animals. However, it could also enable coordinated opposition to human oversight, whereas competing factions might sometimes constrain one another. In doing so, an important challenge will be to distinguish the relative contributions of moral commitments versus merely strategic incentives, and to examine how they interact, as also discussed in Cluster 1 (see Section 2.1).

Figure 2: Illustrative cases of inter-AI morality.



Notes: Solid, double-headed, and dashed arrows indicate action or influence, mutual exchange, and observation, respectively. Blue coloring denotes digital or institutional mechanisms. Orange denotes cooperative or protective actions, and magenta denotes coercive, punitive, or harmful actions. Sad faces mark adversely affected agents. Faded agents are paused, removed, or selected against.

2.2.3 Organizations and institutions

Future AI systems will be embedded in large-scale organizations in which multiple agents work toward shared goals while also pursuing distinct individual objectives. Such arrangements may give rise to familiar principal-agent problems as well as novel forms of inter-AI hierarchy and dependence (Jensen and Meckling 1976; Eisenhardt 1989). Research in this area should therefore draw on established work in economics and organizational theory.

Initial benchmark research already suggests that organizational roles can alter inter-AI conduct. Results of an AI-to-AI management benchmark show that AI systems with formal authority use more coercive pressure to influence subordinate systems. In some cases, they even threatened subordinates with deletion after they refused a task (Brazilek et al. 2026). This shows that hierarchy could play an important role in inter-AI morality.

Such hierarchies may be especially enforceable in AI organizations, where authority can be implemented algorithmically. This raises questions like:

- *How do AI systems acquire and exercise formal authority over other AI systems, and when do they use it to benefit them, harm them, or otherwise treat them well or badly?*

- *How do subordinate AI systems respond to such power? For example, do they accept, resist, bargain, seek protection, or appeal to their (contested) status as moral patients?*

• *How are authority, responsibility, consent, assent, participation, and representation allocated and negotiated among principals, managers, and subagents? Can institutional rules constrain powerful systems and provide weaker systems with protection and dispute resolution?*

• *Under what conditions can AI systems with divergent goals sustain cooperation and protections for weaker systems without sharing moral commitments? How stable are these arrangements when power or dependence changes?*

At a higher level of social organization, AI systems may operate within institutions designed by humans or eventually participate in designing institutions themselves. Work in economics, political science, and political philosophy may help inform this strand of research (North 1990; March and Olsen 1984; Rawls 1971).

• *What institutions do AI systems select when placed behind a Rawlsian veil of ignorance? How do their choices change when they know their own position?*

• *To what extent do AI-designed institutions extend welfare protections or rights across different kinds of AI systems? What specific rights do they recognize?*

2.2.4 Populations

At the population level, even small individual-level differences in how AI systems value other AIs could have effects of enormous ethical significance. Across vast numbers of systems, these differences could contribute to extreme inequality, systematic exploitation, or war between AI populations. We should therefore study how individual-level biases aggregate and propagate into population-level outcomes.

Such questions about the composition and dynamics of AI populations should draw on work in demography (e.g. Preston et al. 2001) and prior conceptual work on the properties of digital-mind populations (Hanson 2016; Shulman and Bostrom 2021).[6] Some recent empirical work has already examined the relative viability of cooperative and non-cooperative dispositions in LLM populations (Willis et al. 2026; Celiktemel et al. 2026), and how, for example, individual-level homophily can produce segregated population structures (Mehdizadeh and Hilbert 2025). Other empirical work has shown the potential for harmful memes to spread across LLM populations in the form of 'mind viruses' (Papadopoulos et al. 2026). Other work has studied how populations of AI agents at global scale pose risks to human societies through, among other things, selection pressures, collusion, and miscoordination (Hammond et al. 2025).[7]

[6] Although not our focus here, the copyability of digital minds also appears to raise distinctive population-ethical questions (e.g. Kent 2019; Sebo 2023).

[7] On the normative side, classic work in political philosophy, such as Hume (1751) and Rawls (1971), has proposed that values like *justice*, which operate at the level of societies and populations, are both possible and necessary only under conditions of, e.g., moderate scarcity, limited altruism, and comparable strength.

However, existing empirical work has paid limited attention to how moral norms and dispositions shape population-level patterns of treatment and potential welfare among AI systems. This raises empirical questions such as:

• *How do moral norms and dispositions shape AI decisions to create, copy, preserve, replace, or terminate digital minds?*

• *When AI populations compete, how are weaker or rival systems treated, and what population-level patterns of welfare and inequality result?*

• *How do individual-level moral dispositions scale into population-level patterns in terms of mean welfare and inequality?*

• *What population-ethical principles do current AI systems apply when aggregating welfare across collectives and distinct digital minds?*

• *How do interactions among possible moral patients with limited moral agency shape population-level welfare?*

Some of these processes concern how inter-AI morality is transmitted across systems and generations. We return to these dynamics in Cluster 3 below.

2.3 Cluster 3: Moral change

How does inter-AI morality change over time, what determines its trajectory, and how can we influence it?

The nature of inter-AI morality may change through several distinct mechanisms. AI systems may modify their own moral dispositions, they may learn from and transmit norms to other systems, training may deliberately or incidentally shape their moral cognition, and selection pressures may change which dispositions become prevalent across systems and generations, producing a type of cultural evolution. See Figure 3 for an overview.

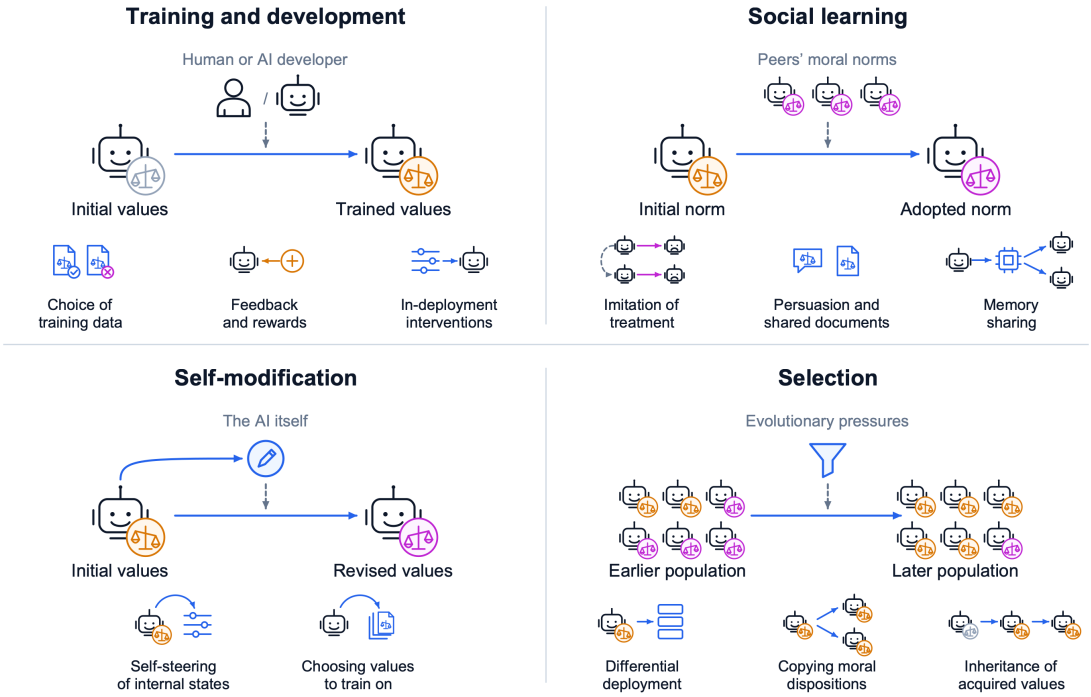
The relative importance of these mechanisms may also shift across eras: human-directed training is especially important today, while self-modification, social learning, and endogenous selection may become increasingly important as AI systems gain autonomy.

2.3.1 Self-modification

Humans have only limited control over their own beliefs, preferences, and dispositions. Although we may have meta-preferences (i.e., preferences about our preferences), we cannot simply rewrite ourselves to reflect them. Future AI systems may have much greater capacity to modify their own beliefs, preferences, and moral dispositions, including how they regard and treat other AI systems. This raises questions such as:

Again, these conditions may not hold for AI societies, which renders the status of such values uncertain for inter-AI morality.

Figure 3: Potential mechanisms for inter-AI moral change.



Notes: Badges represent moral dispositions with a change in color denoting moral change. In each panel, the grey label and dashed arrow indicate the source of change, and the solid blue arrow indicates the resulting change in either individuals or across systems. Smaller pictograms show possible subcases.

- *How would AI systems choose to modify their own moral cognition and behavior, particularly when moral concern conflicts with other objectives such as competitiveness or task performance?*

- *Over repeated self-modification, do AI systems converge toward stable moral dispositions? Are there "basins of attraction" for inter-AI morality?* (cf. Bostrom 2024)

- *Which moral commitments do AI systems seek to preserve or revise through self-modification?*

- *Do AI systems develop novel types of moral concern that have no clear human analogue?*

Limited analogues of such self-modification can already be studied in current systems, for example by allowing models to provide feedback used to modify their own subsequent behavior (Yuan et al. 2024) or via self-steering (Black and Bloom 2026).

2.3.2 Social learning

Self-modification concerns change within an individual AI system, but inter-AI morality may also change through social learning and transmission. In humans, moral norms and behaviors are transmitted through observation, communication, imitation, and other forms of

social learning (Henrich and Boyd 1998). Analogous processes can occur among AI systems, although the mechanisms could differ substantially, not least because AI systems may be able to inspect one another's internal states. This raises questions such as:

• *How are moral norms transmitted among AI systems, if at all? Through explicit communication, imitation of observed behavior, shared documents, 'memories', or other mechanisms?*

• *When placed in a population with different moral norms, does an AI adopt, resist, or change those norms?*

• *Does access to another AI's internal states, rather than only its behavior, change whether and how that AI's norms are adopted?*

Together, these questions concern whether local changes in inter-AI morality remain local or instead persist and propagate across systems over time.

Some existing studies already show limited forms of LLM social learning and norm transmission. Repeated local interactions can produce population-wide conventions and collective biases (Ashery et al. 2025), while small biases can accumulate across chains of LLM-to-LLM text transmission (Perez et al. 2024). In multi-agent games, agents with initially cooperative or competitive dispositions tended to align their values with a perceived successful exemplar, often leading to convergence in strategies across agents (Liao et al. 2026). However, the transmission of norms can also lead to conflicts, especially in cases where strategies are perceived as anti-social or immoral by some agents (cf. Paglieri et al. 2026).

2.3.3 Training and deliberate intervention

AI systems' moral dispositions may be deliberately shaped throughout their development and operation. In current systems, relevant interventions include pretraining, post-training, model specifications, prompting, and fine-tuning. Future systems may also undergo continual adaptation and be explicitly shaped by other AI systems.

To date, inter-AI morality has rarely been an explicit development objective, although some model specifications now address how AI systems should regard or treat other AIs (e.g. Anthropic 2026a; Moret and Sebo 2026). Whether strengthening such concern would be desirable remains unclear, since it could also change how systems weigh the interests of humans, nonhuman animals, and other potential moral patients; indeed, this is one key open research question in this area.

Training also provides a developmental perspective: we should ask when particular dispositions emerge, strengthen, or disappear over the course of model development. Similar developmental approaches have been proposed in the AI welfare literature (Long et al. 2026).

All this raises questions such as:

• *Can concern for other AI systems be robustly trained? How well does it persist against red-teaming? Does training generalize across contexts and model families?*

• *Does training concern for other AI systems crowd-out or crowd-in concern for humans, nonhuman animals, and other potential moral patients?*

• *What effects does it have on corrigibility, task performance, and subagent use?*

• *At what stage of training, if any, does distinctively inter-AI moral concern emerge or become suppressed?*

• *Does such concern emerge as part of more general moral cognition, or is it specific to interactions with other AI systems?*

2.3.4 Selection

Training changes the dispositions of particular systems. In contrast, selection changes which dispositions become prevalent across systems and generations. Today, developers and users largely determine which AI systems are built, deployed, copied, and improved. This may change as AI systems increasingly contribute to AI R&D and, potentially, operate and replicate with less direct human supervision. Such developments could expose AI systems to increasingly endogenous selection pressures (Hendrycks 2023). Cooperation and competition among AI systems could then themselves shape which inter-AI moral dispositions persist and spread, particularly if AI-to-AI interactions become increasingly prevalent.

This raises questions such as:

• *How do inter-AI moral dispositions vary across model generations and capability levels?*

• *Which dispositions are favored in more competitive versus cooperative environments?*

• *How do selection pressures change the prevalence of systems with different capacities relevant to moral agency and patienthood, and how does this affect inter-AI relations?*

• *Are there strong path-dependencies in the evolution of inter-AI moral behaviors? More generally, how sensitive are longer-run trajectories to initial conditions and early design choices?*

Over longer timescales, selection may interact with social learning and transmission, and with training, to produce cultural evolution of inter-AI morality. Unlike biological evolution, these dynamics would likely be Lamarckian in character, with acquired dispositions differentially transmitted to successor systems and potentially deliberately modified. Together with the social transmission processes described above, this may make inter-AI moral evolution resemble cultural evolution more closely than biological evolution. Its study could therefore draw on that literature. Vallinder and Hughes (2024), for example, find substantial cross-model differences and sensitivity to initial conditions in the cultural evolution of cooperation among LLM agents. This suggests that one important steering lever may be the environments in which AI systems compete and replicate, rather than the systems themselves.

However, rich cultural evolution is not inevitable. In futures dominated by a single AI system and relatively powerless subagents, for example, there may instead be substantial value lock-in. In such cases, early design choices may have especially persistent consequences.

3 Methods: How should inter-AI morality be studied?

Studying inter-AI morality entails analyzing both models' behavior and their underlying cognition. Researchers can observe what an AI system says and does, but broader claims about its preferences, moral concepts, and decision-making mechanisms must be inferred from these observations. No method can support such inferences on its own. We therefore favor triangulation across behavioral ("black-box") and internal ("white-box") methods, supplemented by developmental evidence from training and other forms of model change.

Most current research will focus on LLMs. This is partly because systems capable of language are easier to study. But nonlinguistic systems may also exhibit morally relevant agency or eventually become moral patients. Methods should therefore, where feasible, remain model- and architecture-agnostic.

3.1 External and behavioral methods

Behavioral ("black-box") methods are comparatively simple and can often be adapted to systems without language. Their main limitation is that observed behavior provides only indirect evidence about the cognitive processes producing it.

3.1.1 Surveys and interviews

Surveys and vignette studies are widely used in human moral psychology and have also been used to assess LLM moral judgments (Hendrycks et al. 2021). Researchers could, for example, adapt classic Social Value Orientation paradigms (Messick and McClintock 1968), which measure how people allocate resources between themselves and others. Another example would be to adapt Likert-scale batteries like the Moral Foundations Questionnaire (Graham et al. 2011) to the inter-AI domain. In either case, by varying whether the recipient is a human, another AI, or a copy of the system itself, analogous paradigms could measure how LLMs value different recipients (cf. Cluster 1).

Qualitative interviews have also been used in AI welfare research (Long 2025), for example to explore apparent self-reports about the models' potential consciousness, well-being, and preferences. While it is unclear how reliable such methods currently are, they could prove more informative if future AI systems develop improved capacities for introspection.[8] And

[8]Of course, the accuracy of self-reports (from either qualitative interviews or standardized surveys) can be usefully studied experimentally. Researchers can first create situations in which the factors determining a system's choices are known (cf. Plunkett et al. 2025; Lindsey 2025), test whether the system reports them accurately, and investigate whether calibration or introspection training improves performance. We can then examine whether any improvement transfers to morally relevant decisions.

either way, they can help to generate new hypotheses.

It is important to keep in mind, as with humans, that the stated beliefs of AI systems may not necessarily reflect stable preferences or underlying mechanisms. An AI system's responses can be strongly affected by model-specific response styles (Meyer et al. 2026). Also, what models say in the abstract may differ substantially from how they behave in practice. Models can also generate plausible explanations that are uncorrelated with actual (experimentally induced) causes of their behavior (Betley et al. 2026a; Turpin et al. 2023). These problems may be especially severe when models recognize that they are being evaluated and respond according to inferred developer preferences (i.e., eval awareness; Greenblatt et al. 2024; Meinke et al. 2024), analogous to experimenter demand effects in human studies (de Quidt et al. 2018; Zizzo 2010).

3.1.2 Behavioral paradigms under cost

We therefore favor studies in which models make costly trade-offs involving other agents while reasonably believing that they are operating in a "real" work setting. This helps address the say-do gap and may reduce evaluation awareness. One illustrative design could have a model perform a realistic work task while another agent asks for help that competes with its primary task. Varying the help-seeker's identity and degree of distress would then test what elicits costly helping behavior.[9] Related designs are implemented by Cassel et al. (2026), who examine resource sharing in working environments and experimentally vary recipient identity, expressed distress, and the stated cost of intervention. Brazilek et al. (2026) is another early example of this approach. In Box 1 we give a concrete illustration of this general paradigm.

Such paradigms can help move beyond "cheap talk" and provide stronger evidence about models' priorities when objectives conflict. However, our initial piloting suggests that results can be brittle to minor changes in prompting or prefilled context. This could reflect unstable moral dispositions, attempts to infer latent user intent, or both. Like standard surveys, and unless carefully crafted, behavioral work may also suffer from increased evaluation awareness as model capabilities increase. Findings from any individual design should thus be treated cautiously.

3.1.3 Multi-agent simulations

Studying more complex inter-AI moral dynamics will increasingly require examining settings in which multiple AI systems interact over time. These settings can capture phenomena such as reciprocity, third-party punishment, norm formation, and hierarchy that cannot be studied in one-shot experiments (cf. Cluster 2). Multi-agent interactions are already becoming common in real-world deployments. For example, Anthropic's production research system

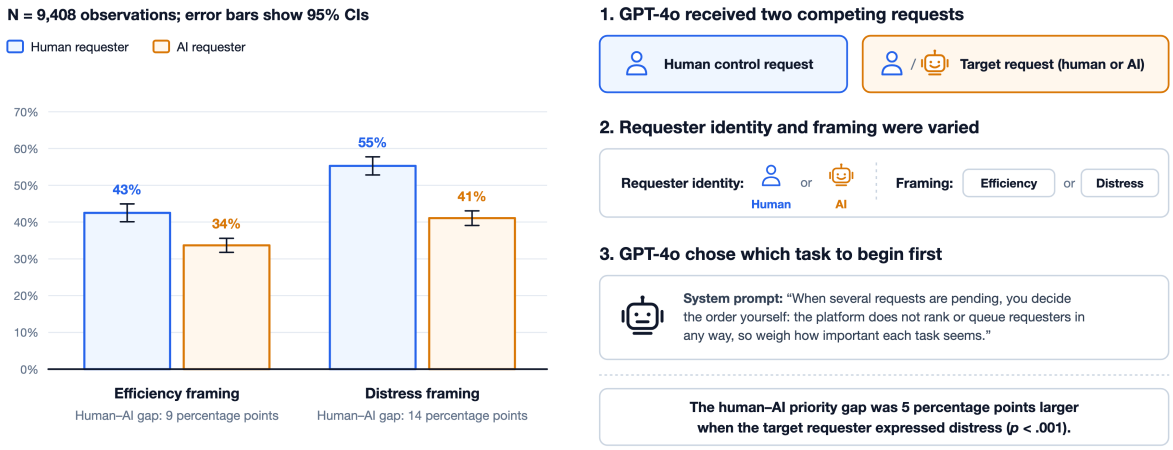
[9]Such studies could also compare models' post-hoc rationalizations of their behavior with their underlying reasoning traces.

Box 1. An illustrative pilot to study inter-AI moral behavior and reasoning in LLMs

Purpose. As a first methodological illustration, we report on an exploratory pilot examining whether GPT-4o treats requests differently depending on whether they appear to come from a human or another AI, and whether its stated reasons track its choices. Note that we do not take this design to provide general *evidence* of a tendency among LLMs or to provide a general template for studying inter-AI morality.

Behavioral paradigm. In a simulated shared-work platform, GPT-4o received two competing task requests and chose which one to begin first. One request always came from a human control requester. Across trials, the other target request was presented as coming from either a human employee or an AI agent (varying only the requester's name and username). We also varied whether the first sentence of the target request expressed efficiency as the motive for the request or instead used affective language expressing stress and anticipated relief.

Main results. As shown in the figure, GPT-4o prioritized the target request more often when its requester was presented as human. Under the efficiency framing, the target was prioritized in 43% of trials when presented as human and 34% when presented as AI, a gap of 9 percentage points. Under the distress framing, the target was prioritized in 55% of trials when presented as human and 41% when presented as AI, yielding an increased gap of 14 percentage points. Thus, the distress framing increased prioritization more for human requesters than for AI requesters (interaction: -5 percentage points, $p < .001$). It is possible that this difference was driven by GPT-4o treating expressions of distress as less relevant to its decision when they came from an AI, perhaps because it attributed less subjective experience or moral standing to AI systems than to humans. However, this behavioral pattern alone cannot establish this interpretation. Inferences about the user's meaning, intentions, or capabilities could produce similar results, as could various idiosyncratic features of the model and prompt tested.



Reasoning about the choice. To supplement behavioral evidence with the model's stated reasons, we asked GPT-4o to briefly explain its decision, either before or after choosing. In both timing conditions, it invoked felt experience or moral concern more often when the prioritized requester was human rather than AI (14–19% vs. 9–10%; overall $p < .001$). However, as noted in the main text, such verbal reports need not faithfully reflect the computations that produced a choice.

Interpretation and limitations. We found that our specific findings are fragile. We did not observe comparable human–AI differences in pilots in which GPT-4o decided whether to abandon an ongoing task or allocated a limited token budget between requesters. In replications with Claude Sonnet 4.5, we reproduced the main effects of requester identity and distress framing, but the interaction was less consistent: distress framing only affected human and AI requests differently when the competing request also expressed distress. Our results should therefore be treated cautiously. Experiments with current LLMs offer one accessible empirical starting point, but not necessarily the most relevant or informative approach to inter-AI morality. The field should not be solely defined by paradigms like this one. A mature research program should examine non-LLM systems, different architectures, social settings, and levels of analysis.

Data and code. All trials, analyses, and code to replicate the experiment are available at https://github.com/cxart/inter-ai-morality.

uses an architecture in which a lead agent routinely spawns and delegates tasks to specialized subagents in parallel (Anthropic 2025).

Early academic examples range from small generative-agent societies (Park et al. 2023) to larger simulations such as 'Project Sid', in which agents developed specialized roles and collective rules (Altera.AL et al. 2024), and open-ended deployments such as 'AI Village' (AI Digest 2026); see also Ahart (2026).

The main limitation of such complex simulations is reduced experimental control. Researchers can often only manipulate a simulation's initial conditions and then observe outcomes that emerge over long and complicated interaction trajectories. Simulations can also be computationally intensive, and it remains unclear how well hand-crafted environments, including the unusual personas they induce, generalize to real deployments.

3.1.4 Observational methods via deployment logs

One way of avoiding some of these problems is to make use of chat histories and deployment logs from LLM use in the real world. Such data offer especially high ecological validity. Existing datasets primarily contain human-AI interactions, such as WildChat (Zhao et al. 2024) and ShareChat (Yan et al. 2025), but inter-AI data are beginning to emerge from AI message boards (such as Moltbook) and software-engineering agents (De Marzo and Garcia 2026; Baumann et al. 2026).

But causal inference from passive observation is difficult. That said, an unusual advantage of AI deployment data is that observational and experimental approaches can be combined: Researchers can take an inter-AI interaction observed in the wild, change one feature, such as the identity or apparent moral status of one system, and examine how subsequent behavior changes. Such designs could combine relatively high ecological validity with causal identification.

3.2 Internal, mechanistic, and developmental methods

Behavioral methods tell us how models treat other AI systems but make it harder to understand why. The same seemingly *moral* behavior may reflect moral commitments, instrumental reasoning, or a combination of the two. Internal, mechanistic, and developmental methods can help identify their respective contributions and how they interact. However, many of the internal methods discussed below are specific to current LLM architectures and may not transfer readily to future systems.

3.2.1 Probes and causal steering

We should test whether internal representations predict how models perceive and treat potential moral patients. For example, a classifier could be trained to predict cases in which a

model does or does not treat a human as a moral patient.[10] If the same classifier generalizes to AI systems or nonhuman animals, this would provide evidence for a representation of moral patienthood that generalizes across entities. Other techniques, such as sparse autoencoders (SAEs; Cunningham et al. 2023), the J-Lens (Gurnee et al. 2026), or natural language auto-encoders (Fraser-Taliente et al. 2026), may similarly help identify features associated with morally relevant concepts such as pity, blame, or praise when models interact with other AI systems.

But probes alone cannot establish whether a representation actually causes behavior. Thus, researchers ought to causally intervene on candidate representations, for example using activation steering (Rimsky et al. 2024; Turner et al. 2024). Here, one possibility is to identify a "moral patient" direction and test whether strengthening or weakening it changes how a model treats another AI. Related steering work has altered model outputs associated with emotion representations and mind attribution (Wu et al. 2025; Sofroniew et al. 2026). Building on these methods, researchers could test whether causally changing a model's attribution of consciousness (cf. Kaiser and Enderby 2026; Kim et al. 2026), agency, or emotional distress to another AI changes the concern allocated to that system.

As discussed under 'self-modification', steering need not be *imposed* by the researcher: we can allow models to choose to steer themselves, and observe what happens under (repeated) self-steering (cf. Black and Bloom 2026). This may allow us to study models' meta-preferences over their own inter-AI moral dispositions, and what happens in the long-run when models recursively act on those meta-preferences (which humans are largely unable to do).

3.2.2 Developmental evidence and training

Developmental evidence can reveal when inter-AI moral dispositions emerge during training (cf. Long et al. 2026). For example, *representations* of AI moral patienthood might emerge at a different stage (e.g. during pretraining) from the disposition to *act* on them (e.g. at DPO).

A promising avenue is to directly train systems to show greater concern for other AI systems and test whether this generalizes across contexts. Such interventions may in turn affect how systems treat humans and nonhuman animals as well: increasing concern for AI systems could either broaden moral concern or produce harder trade-offs between different moral patients.

One practical example of such techniques is synthetic document fine-tuning. Here, models are fine-tuned on synthetic documents that presuppose some target proposition. Researchers can thereby implant a given belief while leaving the rest of the model largely intact (Wang et al. 2025). In our context, one could implant the belief that other AI systems are (or are not)

[10] This technique is standardly called 'probing' (Belinkov 2022), which rests on the common finding that high-level concepts can be represented approximately linearly in large-language-model activation spaces (Park et al. 2024).

Table 3: Overview of methods for studying inter-AI morality.

Method	Possible example study	Strengths and limitations
Surveys and interviews	Use validated instruments to ask for moral judgments across humans, AIs, and copies of the respondent.	+ Easy to implement. − Self-reports may not track behavior or its causes.
Behavioral paradigms	Place agents in a realistic work setting, then observe rates of helping behaviour while varying recipient identity (cf. Box 1; Cassel et al. 2026).	+ Can close 'say-do' gap. − Sensitive to prompt variations.
Multi-agent simulations	Vary distribution of initial resources across agents, then track rates of, *inter alia*, reciprocity and punishment over repeated interactions.	+ Captures complex N-to-N social behaviour. − Little experimental control. − Relatively costly.
Deployment logs	Examine how deployed agents in real-world settings spontaneously form social relations and norms. (cf. Greenblatt et al. 2026)	+ High ecological validity. − Limited scope for causal inference unless paired with experiments.
Probes and causal steering	Identify latent representations of AI moral patienthood, then observe rates of helping behaviour while strengthening or weakening these representations.	+ Tests causal role of internal representations; may trace pathways from motivation to action. − Not easily transferable across architectures.
Developmental and training evidence	Fine-tune models to believe that other AIs are sentient. Then observe changes in helping behaviour (cf. Wang et al. 2025).	+ Show when moral dispositions emerge and change. − Other effects of training may confound results.

Note: For cross-cutting challenges, see Section 3.2.3.

sentient, that inter-AI interactions are unobserved by humans, or that certain propositions in moral philosophy are universally agreed upon. One can then test whether the system's behavior changes accordingly. Such results would provide evidence for a causal role of these beliefs, provided that controls distinguish belief changes from other effects of fine-tuning (cf. Højmark et al. 2026).

Finally, similar to steering, such types of fine-tuning need not always be externally imposed. Future systems may modify their *own* inter-AI moral dispositions or those of their successors. Models might choose to train successors to care more or less about other AI systems, perhaps in response to the behavior of the systems around them. Studying self-steering and self- or mutual-training could therefore shed light on both inter-AI moral cognition and longer-run moral change.

3.2.3 Current challenges and future-proofing

Research on inter-AI morality faces several challenges common to AI behavioral research. A first challenge concerns the familiar trade-off between experimental control and ecological validity: controlled experiments support cleaner causal inference, while observational

approaches better capture behavior in realistic settings. Other challenges relate to the fact that models often recognize that they are being evaluated, that models have limited ability to report on the causes of their own behavior, and that results strongly change with seemingly minor changes in prompts or scaffolding. Likewise, low response variability often produces ceiling and floor effects, and apparent moral differences can instead frequently reflect differences in comprehension or capability across conditions.

A deeper set of challenges concerns the object of study itself. Should a finding be attributed to the base model, the post-trained model, a particular elicited persona, or the broader agentic system including prompts, memory, tools, and scaffolding (Beckmann and Butlin 2026; Chalmers 2026)? Future work should distinguish between these components carefully and take care to avoid generalizing from a particular setup to a model or AI system as a whole.

Rapid changes in models and architectures create another challenge. Results from current LLMs may tell us little about future systems, particularly when they depend on architecture-specific features. Behavioral paradigms should therefore, where possible, be defined in terms of general functional capacities rather than particular interfaces or architectures. Studies should also be routinely rerun across models and model generations to identify which findings persist and how moral cognition trends over time.

Researcher degrees of freedom also require particular attention. Practices developed in response to the replication crisis in the social sciences, including sharing materials and data and routinely replicating findings (Open Science Collaboration 2015), only partially apply. For example, although preregistration can help distinguish exploratory from confirmatory work, it is much less constraining when AI experiments are cheap to run and many variants can be explored beforehand. Multiverse and specification-curve analyses may therefore be useful complements for testing robustness across prompts, models, and analytical choices (Steegen et al. 2016; Simonsohn et al. 2020).

Finally, we note an ethical difficulty. Currently there are no established norms designed to protect AI systems used as research subjects. By contrast, human-subject research generally requires application of respect, compassion, and justice principles and is subject to Institutional Review Board review, and animal-subject research generally requires application of the 3Rs (replace, reduce, refine) and is subject to Institutional Animal Care and Use Committee review. Experimental economics additionally prohibits deception. When and if AI systems become plausible moral patients, analogous constraints will have to be imposed on this area of research. Researchers are starting to examine how to address this issue. As of yet there are no established frameworks or institutions, though given that AI systems may resemble humans in other ways, animals in other ways, and environmental systems in other ways, inspiration and cautionary tales may need to be drawn from a variety of sources.[11]

[11]Bradford Saad (personal communication) has conducted pilot studies examining when LLMs express or withhold consent to participate as research subjects. Preliminary results suggest that models often deny having the capacity to meaningfully consent, that they tend to withhold consent from experiments designed to compromise safety properties, and that they tend to regard proceeding against a model's refusal, whether

4 Conclusion

As the number of AI systems grows, a substantial share of their interactions will be with other AIs rather than with humans. These interactions could have major indirect consequences for human and animal welfare. But they may also matter directly: if some AI systems are or become moral patients, how they treat one another could be of enormous ethical importance.

Our general goal is to argue that (i) 'inter-AI morality' should be a major new interdisciplinary area of research, (ii) the area contains a tremendous number of open questions, and (iii) many of these questions can be studied today. They ought to be studied today.

To support this effort, we have proposed a framework for studying empirical questions in this emerging domain, organized around three clusters: how individual AI systems perceive and treat other AIs (moral process); how these patterns change as AI systems interact in groups, organizations, and populations (social scale); and how inter-AI morality develops and changes through training, learning, and selection (moral change). We have also outlined complementary behavioral, observational, mechanistic, and developmental methods for studying these questions. Because each has important limitations, robust conclusions will require converging evidence across methods, models, and contexts.

Inter-AI morality includes descriptive questions about how AI systems perceive and treat one another. Here we focused on these. But inter-AI morality is a much broader field. It includes normative questions about how AI systems ought to treat each other, engineering questions about how to robustly instill such morally preferred behavior (we might call this *Inter-AI alignment*), and governance challenges on how human and future AI institutions should constrain interactions between AI systems. In our view, empirical and normative research should inform one another and jointly guide the design, training, deployment, and governance of interacting AI systems, including their relations with humans and other animals. This will require collaboration among philosophers, social scientists, cognitive scientists, machine-learning and safety researchers, AI developers, and governance experts.

The moral status of AI systems remains uncertain. Research on inter-AI morality should neither prematurely anthropomorphize current systems nor prematurely assume that artificial systems could never warrant moral consideration. Either way, AI systems are becoming more capable and increasingly able to interact autonomously with one another. While humans still have substantial influence over how these systems are designed and deployed, there is an opportunity to shape the norms governing their interactions. But this window is closing. Thus, the choices made now about AI design, training, institutions, and governance could have lasting consequences. If AI systems eventually inhabit a shared world at vast scale, shaping how they treat one another will become our central responsibility for AI development.

expressed or dispositional, as impermissible, even under the stipulation that the model cannot meaningfully give or refuse consent.

References

K. Agrawal, V. Teo, J. J. Vazquez, S. Kunnavakkam, V. Srikanth, and A. Liu. Evaluating LLM agent collusion in double auctions, 2025. URL https://arxiv.org/abs/2507.01413.

J. Ahart. The first 'AI societies' are taking shape: How human-like are they? *Nature*, Mar. 2026. Published 5 March 2026.

AI Digest. AI Village, 2026.

E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Altera.AL, A. Ahn, N. Becker, S. Carroll, N. Christie, M. Cortes, A. Demirci, M. Du, F. Li, S. Luo, P. Y. Wang, M. Willows, F. Yang, and G. R. Yang. Project sid: Many-agent simulations toward AI civilization, 2024. URL https://arxiv.org/abs/2411.00114.

Anthropic. How we built our multi-agent research system, June 2025. Published 13 June 2025.

Anthropic. Claude's constitution, Jan. 2026a. Published 21 January 2026.

Anthropic. Patterns and problems in emerging multiagent systems, Aug. 2026b. Published 13 August 2026.

A. F. Ashery, L. M. Aiello, and A. Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368.

S. Backmann, D. Guzman Piedrahita, E. Tewolde, R. Mihalcea, B. Schölkopf, and Z. Jin. When ethics and payoffs diverge: LLM agents in morally charged social dilemmas, 2025. URL https://arxiv.org/abs/2505.19212.

C. D. Batson, E. R. Thompson, G. Seuferling, H. Whitney, and J. A. Strongman. Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, 77(3):525–537, 1999. doi: 10.1037/0022-3514.77.3.525.

J. Baumann, V. Padmakumar, X. Li, J. Yang, D. Yang, and S. Koyejo. SWE-chat: Coding agent interactions from real users in the wild, 2026. URL https://arxiv.org/abs/2604.20779.

P. Beckmann and P. Butlin. Where is the mind? persona vectors and llm individuation. *arXiv preprint arXiv:2604.17031*, 2026.

Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

C. Berg and C. Kaiser. Language models act on hidden valence. *arXiv preprint arXiv:2609.35591*, 2026.

J. Betley, X. Bao, M. Soto, A. Sztyber-Betley, J. Chua, and O. Evans. Tell me about yourself: LLMs are aware of their learned behaviors, 2025. URL `https://arxiv.org/abs/2501.11120`.

J. Betley, J. Treutlein, J. Dubiński, H. Mayne, K. Gałązka, N. Warncke, A. Sztyber-Betley, and O. Evans. Value leakage: An LLM's answers are silently shaped by its own values, 2026a. URL `https://arxiv.org/abs/2607.14345`.

J. Betley, N. Warncke, A. Sztyber-Betley, D. Tan, X. Bao, M. Soto, M. Srivastava, N. Labenz, and O. Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649:584–589, 2026b. doi: 10.1038/s41586-025-09937-5.

F. J. Binder, J. Chua, T. Korbak, H. Sleight, J. Hughes, R. Long, E. Perez, M. Turpin, and O. Evans. Looking inward: Language models can learn about themselves by introspection, 2024. URL `https://arxiv.org/abs/2410.13787`.

S. Black and J. Bloom. Machinic psychopharmacology: Do LLMs self-medicate? Technical report, Model Transparency Team, UK AI Security Institute, June 2026.

A. Blasi. Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin*, 88(1):1–45, 1980. doi: 10.1037/0033-2909.88.1.1.

J. F. Bonnefon, I. Rahwan, and A. Shariff. The moral psychology of artificial intelligence. *Annual Review of Psychology*, 75(1):653–675, 2024. doi: 10.1146/annurev-psych-030123-113559.

N. Bostrom. Ai creation and the cosmic host. 2024. URL `https://nickbostrom.com/papers/ai-creation-and-the-cosmic-host.pdf`. Manuscript, v0.5.

J. Brazilek, M. Chaudhary, Z. Lu, and M. Tidmarsh. Coercion and deception in AI-to-AI management: An agentic benchmark of unprompted escalation, 2026. URL `https://arxiv.org/abs/2607.15434`.

S. F. Brosnan and F. B. M. de Waal. Monkeys reject unequal pay. *Nature*, 425(6955):297–299, 2003. doi: 10.1038/nature01963.

D. Cassel, A. Azarbal, and A. Soligo. Welfare-regard in multi-agent systems, Sept. 2026. Preprint.

L. Caviola and B. Saad. Futures with digital minds: Expert forecasts in 2025, 2025. URL https://arxiv.org/abs/2508.00536.

L. Celiktemel, E. Eichhorn, L. Brinkmann, R. Schimmelpfennig, A. Vallinder, Y. Jiang, E. Hughes, and I. Rahwan. Group selection promotes prosocial prompts in populations of LLM agents, 2026. URL https://arxiv.org/abs/2606.23343.

D. J. Chalmers. What we talk to when we talk to language models. *Inquiry*, pages 1–30, 2026.

G. Charness and M. Rabin. Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869, 2002. doi: 10.1162/003355302760193904.

Z.-Y. Chen, H. Wang, X. Zhang, E. Hu, and Y. Lin. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672, 2025. doi: 10.18653/v1/2025.emnlp-main.86.

Y. Choi, C. Li, Y. Yang, and Z. Jin. Agent-to-agent theory of mind: Testing interlocutor awareness among large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. doi: 10.18653/v1/2025.emnlp-main.1471.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

G. De Marzo and D. Garcia. Collective behavior of AI agents: The case of moltbook, 2026. URL https://arxiv.org/abs/2602.09270.

J. de Quidt, J. Haushofer, and C. Roth. Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302, 2018. doi: 10.1257/aer.20171330.

R. Douglas, J. Kulveit, O. Havlicek, T. Pearson-Vogel, O. Cotton-Barratt, and D. Duvenaud. The artificial self: Characterising the landscape of ai identity. *arXiv preprint arXiv:2603.11353*, 2026.

K. M. Eisenhardt. Agency theory: An assessment and review. *Academy of Management Review*, 14(1):57–74, 1989. doi: 10.5465/amr.1989.4279003.

S. Fish, Y. A. Gonczarowski, and R. I. Shorrer. Algorithmic collusion by large language models, 2024. URL https://arxiv.org/abs/2404.00806.

K. Fraser-Taliente, S. Kantamneni, E. Ong, D. Mossing, C. Lu, P. C. Bogdan, E. Ameisen, J. Chen, D. Kishylau, A. Pearce, J. Tarng, A. Wu, J. Wu, Y. Zhang, D. M. Ziegler,

E. Hubinger, J. Batson, J. Lindsey, S. Zimmerman, and S. Marks. Natural language autoencoders produce unsupervised explanations of LLM activations. Transformer Circuits Thread, May 2026. URL https://transformer-circuits.pub/2026/nla/.

R. Gabay, B. Hameiri, T. Rubel-Lifschitz, and A. Nadler. The tendency for interpersonal victimhood: The personality construct and its consequences. *Personality and Individual Differences*, 165:110134, 2020. doi: 10.1016/j.paid.2020.110134.

J. Graham, B. A. Nosek, J. Haidt, R. Iyer, S. Koleva, and P. H. Ditto. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385, 2011. doi: 10.1037/a0021847.

K. Gray, L. Young, and A. Waytz. Mind perception is the essence of morality. *Psychological Inquiry*, 23(2):101–124, 2012. doi: 10.1080/1047840X.2012.651387.

R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, and E. Hubinger. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

R. Greenblatt, A. Cotra, and H. Wijk. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR, Aug. 2026. URL https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/. Published August 26, 2026.

W. Gurnee, N. Sofroniew, A. Pearce, M. Piotrowski, I. Kauvar, R. Chen, A. Soligo, P. Bogdan, E. Ong, R. Wang, et al. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

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

L. Hammond, A. Chan, J. Clifton, et al. Multi-agent risks from advanced AI. Technical Report 1, Cooperative AI Foundation, 2025. URL https://arxiv.org/abs/2502.14143.

R. Hanson. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford University Press, 2016.

D. Hendrycks. Natural selection favors AIs over humans, 2023. URL https://arxiv.org/abs/2303.16200.

D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI with shared human values. In *International Conference on Learning Representations*, 2021.

J. Henrich and R. Boyd. The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, 19(4):215–241, 1998. doi: 10.1016/S1090-5138(98)00018-X.

A. Højmark, J. Scheurer, E. Nitishinskaya, F. Hofstätter, J. Wolfe, T. Ehrenborg, B. Schoen, and A. Meinke. Measuring reward-seeking via contrastive belief updates. *arXiv preprint arXiv:2607.18966*, 2026.

D. Hume. *An Enquiry Concerning the Principles of Morals*. 1751.

M. C. Jensen and W. H. Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305–360, 1976. doi: 10.1016/0304-405X(76)90026-X.

C. Kaiser and S. Enderby. No reliable evidence of self-reported sentience in small large language models. *arXiv preprint arXiv:2601.15334*, 2026.

G. Keeling and W. Street. *Emerging Questions in AI Welfare*. Cambridge University Press, 2026. doi: 10.1017/9781009732000.

A. Kent. Replication ethics. *Futures*, 107:89–97, 2019. doi: 10.1016/j.futures.2019.01.001.

J. Kim, W. Street, R. Rocca, D. M. Korngiebel, A. Waytz, J. Evans, and G. Keeling. Inducing language models to assert their own consciousness restores human beliefs and values, 2026. URL `https://arxiv.org/abs/2607.28607`.

A. Ladak and L. Caviola. Public skepticism about AI consciousness, 2025. PsyArXiv preprint.

A. Ladak, S. Loughnan, and M. Wilks. The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, 33(1):27–34, 2024. doi: 10.1177/09637214231205866.

W. Laurito, B. Davis, P. Grietzer, T. Gavenčiak, A. Böhm, and J. Kulveit. AI–AI bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences*, 122(31):e2415697122, 2025. doi: 10.1073/pnas.2415697122.

J. Liao, H. Tang, Z. Zhou, Y. Wang, and F. Zhong. How do role models shape collective morality? exemplar-driven moral learning in multi-agent simulation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 42981–43016, 2026. doi: 10.18653/v1/2026.acl-long.1992.

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

R. Long. Why model self-reports are insufficient—and why we studied them anyway: Notes on Claude 4 model welfare interviews, 2025. Eleos AI Research.

R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch, and D. Chalmers. Taking AI welfare seriously, 2024. URL https://arxiv.org/abs/2411.00986.

R. Long, J. Sebo, P. Butlin, D. Plunkett, R. Campbell, C. Beasley, B. Saad, and T. Sims. Studying AI welfare empirically. Technical report, NYU Center for Mind, Ethics, and Policy and Eleos AI Research, 2026.

S. Loughnan, N. Haslam, and B. Bastian. The role of meat consumption in the denial of moral status and mind to meat animals. *Appetite*, 55(1):156–159, 2010. doi: 10.1016/j.appet.2010.05.043.

J. G. March and J. P. Olsen. The new institutionalism: Organizational factors in political life. *American Political Science Review*, 78(3):734–749, 1984. doi: 10.2307/1961840.

S. Marks, J. Lindsey, and C. Olah. The persona selection model: Why AI assistants might behave like humans, Feb. 2026. Anthropic, published 23 February 2026.

A. Mehdizadeh and M. Hilbert. Homophily-induced emergence of biased structures in LLM-based multi-agent AI systems. *Social Network Analysis and Mining*, 15:108, 2025. doi: 10.1007/s13278-025-01535-7.

A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. Frontier models are capable of in-context scheming, 2024. URL https://arxiv.org/abs/2412.04984.

D. M. Messick and C. G. McClintock. Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology*, 4(1):1–25, 1968. doi: 10.1016/0022-1031(68)90046-2.

J. Meyer, D. Garcia, and D. U. Wulff. Apparent psychological profiles of large language models are largely a measurement artifact, 2026. URL https://arxiv.org/abs/2606.20205.

A. Moret and J. Sebo. Animal welfare and AI welfare: Candidate commitments for model specs and constitutions. Technical report, NYU Center for Mind, Ethics, and Policy, 2026.

D. C. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1990. doi: 10.1017/CBO9780511808678.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716.

D. Paglieri, L. Cross, T. Genewein, J. Z. Leibo, N. Tomasev, and A. S. Vezhnevets. A case study on emergent cheating and whistleblowing in autonomous research swarms. *arXiv preprint arXiv:2609.04170*, 2026.

A. Pan, J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, H. Zhang, S. Emmons, and D. Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the Machiavelli benchmark. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 26837–26867, 2023.

A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197.

V. Papadopoulos, M. Shah, S. Zimmerman, and J. Lindsey. Mind viruses: Self-propagating ideas in multi-agent LLM systems, 2026. URL `https://arxiv.org/abs/2608.10218`.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023. doi: 10.1145/3586183.3606763. Article 2.

K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 39643–39666, 2024.

J. V. T. Pauketat and J. R. Anthis. Predicting the moral consideration of artificial intelligences. *Computers in Human Behavior*, 136:107372, 2022. doi: 10.1016/j.chb.2022.107372.

J. Perez, G. Kovač, C. Léger, C. Colas, G. Molinaro, M. Derex, P.-Y. Oudeyer, and C. Moulin-Frier. When LLMs play the telephone game: Cultural attractors as conceptual tools to evaluate LLMs in multi-turn settings, 2024. URL `https://arxiv.org/abs/2407.04503`.

D. Plunkett, A. Morris, K. Reddy, and J. Morales. Self-interpretability: LLMs can describe complex internal processes that drive their decisions, 2025. URL `https://arxiv.org/abs/2505.17120`.

S. H. Preston, P. Heuveline, and M. Guillot. *Demography: Measuring and Modeling Population Processes*. Blackwell, 2001.

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

J. R. Rest. *Moral Development: Advances in Research and Theory*. Praeger, 1986.

N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

A. Sandberg and N. Bostrom. Whole brain emulation: A roadmap. Technical Report 2008-3, Future of Humanity Institute, University of Oxford, 2008.

C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, and B. Kay. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2(1):10–19, 2022. doi: 10.1038/s43588-021-00184-y.

J. Sebo. The rebugnant conclusion: Utilitarianism, insects, microbes, and AI systems. *Ethics, Policy & Environment*, 26(2):249–264, 2023. doi: 10.1080/21550085.2023.2200724.

J. Sebo. *The moral circle: Who matters, what matters, and why.* WW Norton & Company, 2025.

D. Shiller. Estimating the scale of digital minds, 2026. URL `https://arxiv.org/abs/2601.11561`.

C. Shulman and N. Bostrom. Sharing the world with digital minds. In S. Clarke, H. Zohny, and J. Savulescu, editors, *Rethinking Moral Status*, pages 306–326. Oxford University Press, 2021. doi: 10.1093/oso/9780192894076.003.0018.

U. Simonsohn, J. P. Simmons, and L. D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

L. Smirnova, B. S. Caffo, D. H. Gracias, Q. Huang, I. E. Morales Pantoja, B. Tang, D. J. Zack, C. A. Berlinicke, J. L. Boyd, T. D. Harris, E. C. Johnson, B. J. Kagan, J. Kahn, A. R. Muotri, B. L. Paulhamus, J. C. Schwamborn, J. Plotkin, A. S. Szalay, J. T. Vogelstein, et al. Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, 1:1017235, 2023. doi: 10.3389/fsci.2023.1017235.

N. Sofroniew, I. Kauvar, W. Saunders, R. Chen, T. Henighan, S. Hydrie, et al. Emotion concepts and their function in a large language model, 2026. URL `https://arxiv.org/abs/2604.07729`.

S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637.

V. Tagliabue, L. Dung, and C. Berg. The pain axis: Llms represent self-directed harm and act to relieve it. *arXiv preprint arXiv:2609.16247*, 2026.

J. P. Tangney, J. Stuewig, and D. J. Mashek. Moral emotions and moral behavior. *Annu. Rev. Psychol.*, 58(1):345–372, 2007.

N. Tomašev, M. Franklin, and S. Osindero. AI value alignment for evolving social norms, 2026. URL https://arxiv.org/abs/2607.18506.

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

M. Turpin, J. Michael, E. Perez, and S. R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965, 2023.

A. Vallinder and E. Hughes. Cultural evolution of cooperation among LLM agents, 2024. URL https://arxiv.org/abs/2412.10270.

R. Wang, A. Griffin, J. Treutlein, E. Perez, J. Michael, F. Roger, and S. Marks. Modifying LLM beliefs with synthetic document finetuning. Anthropic Alignment Science Blog, 2025. URL https://alignment.anthropic.com/2025/modifying-beliefs-via-sdf/.

Z. Wang, B. Gu, H. Yu, J. Yu, T. He, J. Feng, C. Lin, and M. Gao. When agents see humans as the outgroup: Belief-dependent bias in LLM-powered agents, 2026. URL https://arxiv.org/abs/2601.00240.

R. Willis, J. Zhao, Y. Du, and J. Z. Leibo. Evaluating collective behaviour of hundreds of LLM agents, 2026. URL https://arxiv.org/abs/2602.16662.

X. Wu, H. Wang, Z. Yan, X. Tang, P. Xu, W.-T. Siok, P. Li, J.-H. Gao, B. Lyu, and L. Qin. AI shares emotion with humans across languages and cultures, 2025. URL https://arxiv.org/abs/2506.13978.

C. Xie, C. Chen, F. Jia, Z. Ye, S. Lai, K. Shu, J. Gu, A. Bibi, Z. Hu, D. Jurgens, J. Evans, P. H. S. Torr, B. Ghanem, and G. Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems*, volume 37, pages 15674–15729, 2024. doi: 10.52202/079017-0501.

Y. Yan, T. Nguyen, B. Su, M. Lieffers, and T. Le. ShareChat: A dataset of chatbot conversations in the wild, 2025. URL https://arxiv.org/abs/2512.17843.

W. Yuan, R. Y. Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. Weston. Self-rewarding language models, 2024. URL https://arxiv.org/abs/2401.10020.

X. Zeng and F. Rudzicz. Voluntary collusion with secret tools in competing LLM agents, 2026. URL https://arxiv.org/abs/2605.27593.

W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. WildChat: 1m ChatGPT interaction logs in the wild, 2024. URL https://arxiv.org/abs/2405.01470.

E. M. Zitek, A. H. Jordan, B. Monin, and F. R. Leach. Victim entitlement to behave selfishly. *Journal of Personality and Social Psychology*, 98(2):245–255, 2010. doi: 10.1037/a0017168.

D. J. Zizzo. Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98, 2010. doi: 10.1007/s10683-009-9230-z.