What is a Morally Aligned AI Agent? Philosophy as a Bridge to Operationalization

Syed Ali Haider [1]

Abstract

Alignment research has a vocabulary problem. We speak of making models reliable, safe, and trustworthy without asking what those words require of a system philosophically. The result is a field that optimizes with great precision for targets it has not clearly defined. This paper argues that philosophy can fix that. Building on the eight-dimension reconstruction of agency developed by Singh (2026), spanning causality, initiative, responsibility, intention, delegation, utility, stability, and computation, we identify three dimensions of morality, normativity, deliberation, and judgment, shared across virtue, deontological, and consequentialist traditions. Crossing the two yields an 8×3 matrix that serves as a coverage map for the alignment field. Our central claim is that current techniques are not randomly distributed across this matrix: we are strong on normativity, increasingly competent on judgment, and consistently weak on deliberation. Making this structure explicit shows where the gaps are, why techniques that succeed on one dimension can fail silently on another, and why treating philosophy as primary rather than decorative is not an aesthetic preference but an engineering necessity.

1. Introduction

There is a version of the alignment problem that gets discussed very little, which is the conceptual version. Not: how do we train models to behave correctly? But rather: what does "correctly" mean, and how would we know if we had achieved it? These are philosophical questions, and the field has largely treated them as someone else's problem. The assumption is that philosophy is vague and alignment is concrete, and we should stay on the concrete side.

[1]Department of Computer Science, Dartmouth College, Hanover, NH, USA. Correspondence to: Syed Ali Haider <syed.ali.haider.gr@dartmouth.edu>.

Preprint. May 17, 2026.

This assumption is wrong, and it is costing us. When a technique like Contrast-Consistent Search (Burns et al., 2024) claims to identify model knowledge, it is implicitly committed to a theory of what knowledge is. When Constitutional AI (Bai et al., 2022) claims to instill values, it is implicitly committed to a theory of what values are. When behavioral evaluations claim to assess alignment, they are implicitly committed to a theory of what alignment means. The philosophically uninformed versions of these commitments are typically thin and inconsistent. The result is that techniques appear to work in one experimental setting and fail to transfer in ways that seem mysterious until you notice that the underlying concept was never clearly specified.

We propose a simple remedy. Philosophy has spent a long time thinking about agents, morality, and what it would mean to be a moral agent. The answers it has reached are not unanimous, but they are structured, and structure is what the alignment field currently lacks. In what follows, we first reconstruct the philosophical definition as synthesized by Singh of an agent across seven dimensions that span the tradition from Aristotle to Friston, then extend to an eighth dimension, computation, that marks the transition from agent in general to AI agent in particular. We then identify the three dimensions of morality that all major ethical frameworks share. We combine these into an organizing framework and show how current alignment techniques map onto it. The conclusion is not that philosophy solves alignment, but that alignment without philosophy will keep solving the wrong things very well.

The Argument in Brief

An agent, philosophically speaking, is not simply a system that acts. Causality is necessary but not sufficient. The tradition adds initiative, responsibility, intention, delegation, utility, and stability, each refining what it means to be a genuine source of action rather than a sophisticated relay. We then add an eighth dimension, computation, to mark what is specific about AI agents: they are sequential decision-makers that perceive, reason, and act through tools under uncertainty, with a language model serving as the policy. Morality, across the virtue, deontological, and consequentialist traditions, requires normativity (the agent

has principles), deliberation (the agent reasons about them), and judgment (the agent or an evaluator can assess conformity). A moral AI agent must satisfy all of these constraints simultaneously. Current alignment research is rich on normativity, thin on deliberation, and inconsistent on judgment. The 8×3 matrix makes this visible; making it visible is the first step toward fixing it.

2. What Is an Agent? A Philosophical Reconstruction

We follow the eight-dimension reconstruction of agency developed in Singh (2026), which traces how the philosophical tradition builds up the concept of an agent in layers, each adding a constraint that the previous left open, and then extends the tradition with one further dimension specific to AI. We work through all eight in order.

The minimum is causality. An agent is a force capable of acting on matter, a primary source of change in the world (Aristotle, 1999). A hammer causes a nail to move, but we do not call the hammer an agent because the force originates elsewhere. The agent must initiate, not merely transmit, which is what Aristotle means when he calls the agent "the first mover that is itself unmoved." This is the dimension of initiative, and it already excludes most reactive systems. A thermostat causes temperature changes but responds entirely to external input. It has causality without initiative.

Moral philosophy adds a third layer. A moral agent is one who "has the ability to discern right from wrong and to be held accountable" (Aristotle, 350). Accountability requires traceability: the agent's actions must be traceable back to something we can call its commitments. This is responsibility, and without it there is nothing to align, because alignment is precisely the project of ensuring that a system's commitments are the right ones.

Anscombe and Davidson formalize the next step. A genuine agent acts for reasons, and those reasons are mental states, specifically desires, beliefs, and intentions, that stand in rational relations to what the agent does (Anscombe, 1957; Davidson, 1963). An agent that says it believes X but whose behavior does not change when X is false does not believe X in any sense the tradition would recognize. This is intention, and it matters because it is the dimension that distinguishes deliberation from pattern-matching.

The legal and commercial tradition adds delegation. An agent is authorized by a principal to act on the principal's behalf (Tiffany, 1903). This relation is not incidental to AI. Every deployed language model is an agent in this precise sense: deployed by an institution, for users, in a relationship structured by authorization and accountability. The principal-agent problem, which asks how principals can ensure agents act on their behalf rather than their own, is not a metaphor for alignment. It is alignment.

Economics adds utility. The rational agent maximizes expected utility, constrained by bounded rationality (von Neumann & Morgenstern, 1944; Simon, 1955). We write this as:

$$a^* = \arg\max_a \; E[U(a) \mid \text{info}] \tag{1}$$

This formalization, whatever its limitations, adds something important: the agent has consistent preferences, and its behavior is a coherent function of them. An agent without consistent preferences cannot be aligned, because there is no stable target for alignment to hit.

Finally, control theory and modern inference add stability. Friston argues that biological agents minimize surprise through active inference (Friston et al., 2010). For AI systems, the analog is behavioral coherence under perturbation: does the agent's behavior remain consistent across paraphrase, distribution shift, and counterfactual pressure? Stability matters because a model that gives aligned answers on the evaluation set but drifts under novel conditions is not a stable moral agent; it is a system that has memorized the right answers for a particular context.

These seven dimensions together constitute what the tradition means by agency:

$$\textbf{Agency} \supseteq \left\{ \begin{array}{l} \textit{causality},\ \textit{initiative},\ \textit{responsibility}, \\ \textit{intention},\ \textit{delegation},\ \textit{utility},\ \textit{stability} \end{array} \right\}$$

To extend this to *AI agents* specifically, Singh (2026) adds one further dimension: **computation**. On this view, an AI agent is a computational system that perceives an environment, reasons, and takes actions through tools to pursue goals on behalf of a principal, sequentially, under uncertainty. Formally, following the standard sequential-decision framing as adopted in Singh (2026), at each step the agent observes $o_t = g(s_t)$, selects $a_t \sim \pi(s_t)$, the environment transitions $s_{t+1} \sim P(s_{t+1} \mid s_t, a_t)$, and it receives reward $R_t(s_t, a_t)$, seeking the policy π^* that maximizes expected return. In modern LLM agents using the ReAct architecture (Yao et al., 2023), the language model *is* π, interleaving reasoning steps, tool calls, and observations in a loop — a framing that Singh (2026) uses to bridge the philosophical reconstruction to current systems.

Adding computation is not a trivial extension. It introduces constraints the philosophical tradition did not anticipate and raises open questions that matter for alignment directly. First, computation makes the agent's reasoning potentially inspectable in a way that human deliberation is not: we can ablate, patch, and probe intermediate states. Whether what we find in those states corresponds to the philosophical notion of intention, the causal commitment to a belief,

remains genuinely unclear (Lanham et al., 2023). Second, computation bounds rationality in a specific way: the agent's deliberation is limited not by cognitive fatigue or attention but by context length, inference budget, and the geometry of the representation space. These are not the bounds Simon had in mind (Simon, 1955), and it is not obvious that the philosophical concept of practical wisdom translates cleanly across them. Third, computation means the agent's policy is a learned artifact, shaped by training data, loss functions, and fine-tuning rather than by lived experience. What it means for such a system to have internalized a principle, rather than memorized its surface form, is the question the deliberation column of our matrix is asking.

$$\textbf{AI Agent} \supseteq \left\{ \begin{array}{l} causality,\ initiative,\ responsibility, \\ intention,\ delegation,\ utility, \\ stability,\ \textbf{computation} \end{array} \right\}$$

3. What Is Morality? Three Frameworks, One Structure

Morality is the system of principles concerning what is right and wrong and the standards by which agents judge their actions and are judged in turn. The three canonical frameworks disagree sharply on where morality lives, but they converge on its structure.

Virtue ethics locates morality in character. A good action is one that a virtuous agent, exercising practical wisdom, would perform (Aristotle, 350). The agent must internalize moral principles deeply enough that they generate the right behavior without explicit deliberation. Deontology locates morality in duty. An action is moral if it conforms to a principle that can be universalized: "act only according to that maxim by which you can at the same time will that it should become a universal law" (Kant, 1785). Consequentialism locates morality in outcomes. An action is moral if it maximizes aggregate well-being, and the agent must track consequences carefully enough to optimize for them (Mill, 1863).

Despite their disagreements, all three frameworks require three things of a moral agent. The agent must have principles, rules, or values that constrain action; this is normativity. The agent must reason about how those principles apply to specific cases; this is deliberation, and it takes different forms in each tradition, practical wisdom in virtue ethics, subsumption under the categorical imperative in deontology, and expected-value calculation in consequentialism. Finally, the agent or an external evaluator must be able to assess whether the action conformed to the standard; this is judgment, and without it there is no accountability.

$$\text{Morality} \supseteq \{\text{normativity},\ \text{deliberation},\ \text{judgment}\} \quad (2)$$

This three-part structure is not a contested philosophical claim. It is the minimal shared commitment of otherwise incompatible moral theories. Any AI system that fails to exhibit all three is not a moral agent in any tradition's sense.

4. A Moral AI Agent: The Coverage Matrix

A moral AI agent must satisfy both constraints: it must be an agent in the eight-dimensional sense developed above, and it must be moral in the three-dimensional sense shared across traditions. The intersection of these two constraints generates an 8×3 matrix where each row is a dimension of agency and each column is a dimension of morality.

Agency dim.	Normativity	Deliberation	Judgment
Causality	○	○	○
Initiative	○	○	○
Responsibility	●	○	●
Intention	●	○	●
Delegation	●	○	○
Utility	●	○	○
Stability	○	○	●
Computation	●	○	○

Table 1. The 8×3 coverage matrix. ● = current alignment techniques have meaningful coverage; ○ = sparse or absent

Coverage examples. Intention crossed with normativity describes Constitutional AI: the model is trained to follow written principles that are intended to drive behavior. Intention crossed with deliberation describes chain-of-thought reasoning: the model is asked to reason through its principles before acting. Intention crossed with judgment describes activation patching or steering vector experiments: the researcher tests whether an internal representation actually causes the behavior it is supposed to explain

Reading the alignment literature through this matrix reveals a pattern that is not random. Normativity is the most developed column: RLHF, Constitutional AI, and related techniques have grown sophisticated at instilling principles. Judgment is growing: adversarial evaluations, behavioral benchmarks, and mechanistic interpretability methods are increasingly able to assess whether behavior matches stated principles. Deliberation is the weakest column by a considerable distance. Most current techniques do not require the model to reason through moral principles at all; they impose them as side effects of training, which means the model may produce compliant outputs without anything that functions as deliberation in the philosophical sense.

Why This Matrix Matters

The matrix makes several previously implicit problems explicit. First, it shows why a technique can succeed in evaluation and fail in deployment: the evaluation may cover one cell while deployment exposes a different one. A model with strong normativity but weak deliberation will give correct answers on familiar cases and generate incorrect ones when a novel situation requires reasoning from first principles rather than applying a memorized rule.

Second, it clarifies the belief-attribution problem. Ramsey's functionalism holds that what makes something a belief is not its physical form but its functional role: does it guide action under stakes, and does it survive perturbation (Ramsey, 1931)? A model that produces chain-of-thought reasoning but whose outputs do not change when that reasoning is ablated does not believe what it says. Belief attribution therefore requires at minimum three things: an intervention demonstrating that the representation causally drives behavior, perturbation tests confirming the effect persists across paraphrase and distribution shift, and multiple independent operationalizations, verbal, behavioral, and mechanistic, that converge. When they diverge, the divergence is not noise. It is diagnostic of which dimensions are and are not satisfied.

Third, the matrix explains a puzzling feature of the field's structure. Safety research, interpretability research, and evaluation research have been developing in parallel with relatively little exchange, despite obviously overlapping goals. The matrix shows why: they are each working on different columns. Safety research is principally concerned with normativity. Interpretability research is principally concerned with judgement and deliberation, asking how representations drive behavior. Evaluation research is principally concerned with judgment. The field is not fragmented because its practitioners have failed to communicate; it is fragmented because the underlying conceptual structure has not been made explicit enough to show that these efforts belong to a unified constraint.

5. Related Work

The philosophical dimensions we employ are not new. The principal-agent problem has been formalized extensively in economics (Ross, 1973; Meckling & Jensen, 1976). Proposals for value alignment similarly invoke delegation and utility (Russell, 2022). Moral agency in AI has been discussed in applied ethics (Floridi & Sanders, 2004; Wallach, 2008), though typically at a level of abstraction that does not reach operationalization. What this paper contributes is the explicit mapping from philosophical structure to alignment technique, using the 8×3 matrix as the organizing device, and the specific identification of deliberation as the underserved dimension.

6. Discussion

We have argued that the alignment field lacks a coherent conceptual vocabulary and that philosophy supplies one. The argument has a limitation worth acknowledging directly: a coverage map is not a solution. Identifying that the deliberation column is weak does not immediately tell us how to fill it. Chain-of-thought reasoning is a partial answer, but its faithfulness to actual model computation is disputed (Lanham et al., 2023), and techniques that verify rather than merely elicit deliberation do not yet exist at scale.

A second limitation is that our philosophical reconstruction draws on a tradition that was not designed for computational systems. The classical concept of practical wisdom, for instance, presupposes something like embodied engagement with the world that a language model may not have. Adapting these concepts to AI is not a mechanical translation; it involves substantive philosophical work that goes beyond what a short paper can do.

What we do claim is that the matrix is useful even without these answers. A coverage map that shows which cells are filled tells the field where to invest next. A conceptual vocabulary that is shared across safety, interpretability, and evaluation research creates the conditions for these subfields to recognize that they are working on the same problem. And a philosophical definition of deliberation, however imperfect its computational analog, sets a target that the field can try to hit rather than optimizing indefinitely for a concept it has not defined.

7. Conclusion

The alignment problem is partly an engineering problem and partly a conceptual one. The engineering part receives enormous attention. The conceptual part, what exactly we are trying to align and what it would mean to have achieved it, receives much less. This paper has argued that the philosophical tradition offers the structure the conceptual part requires. Eight dimensions of agency, three dimensions of morality, and a matrix that maps their intersection onto the alignment literature gives the field a vocabulary precise enough to identify gaps and priorities that the current vocabulary obscures. Philosophy does not solve alignment. But alignment without philosophy will keep solving problems that are adjacent to the real one.

References

Anscombe, G. *Intention*. Oxford University Press, 1957.

Aristotle. *Physics*. Oxford University Press, 1999.

Aristotle. *Nicomachean Ethics*. 350. Translated by W.D. Ross.

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

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision, 2024. URL https://arxiv.org/abs/2212.03827.

Davidson, D. Actions, reasons and causes. *Journal of Philosophy*, 60(23):685–700, 1963.

Floridi, L. and Sanders, J. W. On the morality of artificial agents. *Minds and machines*, 14(3):349–379, 2004.

Friston, K. J., Stephan, K. E., Montague, R., and Dolan, R. J. Computational psychiatry: the brain as a generative model. *Current Opinion in Behavioral Sciences*, 25:85–93, 2010.

Kant, I. *Groundwork of the Metaphysics of Morals*. 1785.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C. E., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T. J., Maxwell, T. D., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Meckling, W. H. and Jensen, M. C. Theory of the firm. *Managerial behavior, agency costs and ownership structure*, 3(4):305–360, 1976.

Mill, J. S. *Utilitarianism*. 1863.

Ramsey, F. P. Truth and probability. *The Foundations of Mathematics and other Logical Essays*, 1931. Written 1926, posthumously published 1931, edited by R.B. Braithwaite, Routledge & Kegan Paul, London.

Ross, S. A. The economic theory of agency: The principal's problem. *American Economic Review*, 63(2):134–139, 1973.

Russell, S. Human-compatible artificial intelligence. *Human-like machine intelligence*, 1:3–22, 2022.

Simon, H. A. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1):99–118, 1955.

Singh, N. Ai agents course. Course lecture, Dartmouth College, 2026. Winter 2026.

Tiffany, J. On the right of agency. *International Journal of Ethics*, 13(3):335–353, 1903.

von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Wallach, W. Implementing moral decision making faculties in computers and robots. *Ai & Society*, 22(4):463–475, 2008.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. arXiv:2210.03629.