Traversing the Uncanny Valley

Artificial Wisdom & the Limits of Alignment

Marc A. Moffett

Penultimate Draft—Do not cite without permission

ABSTRACT: AI alignment is commonly framed as the problem of constraining artificial systems to conform to human values or moral judgments. I argue that this framing rests on a mistaken picture of moral competence. General moral principles face a guidance problem: if simple enough to guide judgment, they omit context-sensitive moral nuance; if sufficiently qualified to capture that structure, their application presupposes the very competence they were meant to supply. I develop an alternative account of moral competence as a form of understanding. Building on a conception-based theory of understanding, I introduce a mechanism for conceptual productivity, whereby an integrated but incomplete conception can place an agent under pressure to acquire new concepts or appropriately anchor previously acquired ones. Applied to the practical domain, this yields a form of (wisdom-based) moral particularism and a corresponding ideal of artificial wisdom: integrated, action-guiding understanding in which evaluative concepts are anchored within a wider conception of value, agency, reasons, and the world. I argue that neither the absence of phenomenal consciousness nor the orthogonality of instrumental intelligence and goals precludes such wisdom, and that objective reasons can constrain artificial judgment without requiring convergence with human verdicts. The central safety risk is therefore a normative uncanny valley: a capability–wisdom gap in which systems acquire powerful practical agency and the appearance of moral competence before acquiring the understanding needed to control its application. Alignment may be instrumental scaffolding, but it should not be mistaken for the final normative target.

1. The Alignment Problem and Its Moral Assumptions

A perennial worry about genuinely intelligent artificial systems is that, unless they are given ethical constraints, they will be left morally adrift. A system may be exceptionally capable with respect to prediction, planning, and control while having no concern for whether its activities cause harm.[1] The natural response to this concern is some form of moral alignment. The system must be in some way constrained to make judgments that align with our moral judgments, rather than being contrary to them.[2] The proposals for achieving alignment differ, but they share a basic feature, namely, human beings supply, directly or indirectly, the normative framework within which the artificial system is constrained to act.

This framing is understandable. Systems with substantial causal power but no sensitivity to harm would plainly be dangerous. But the framework presupposes a substantive theory of moral understanding. Because of this, alignment is primarily treated as an engineering problem concerning how to reproduce or stabilize desirable outputs. But before asking how to implement moral competence artificially, we first need to consider what it consists in. If moral competence is fundamentally about conformity to general principles, then alignment may be a difficult but coherent project. We must identify the right principles, find a way to utilize them adequately, and ensure that the

[1] Bostrom's (2014) thesis of orthogonality is the canonical statement. I will return to Bostrom's discussion in section 6.2.

[2] Alignment is often put in anthropic terms so that alignment is with human preferences or human flourishing. Philosophically, this is unduly parochial and the ethical framing is more satisfactory.

system accords with them. If, however, moral judgment is not fundamentally rule-governed, then a central assumption of the alignment project is mistaken.

The worry here is not simply the existence of moral disagreement. Disagreement creates an obvious difficulty concerning whose values should be encoded. But the problem I wish to raise is independent of this concern. It is that even our best ethical theories appear to misdescribe the competence exercised in good moral judgment. Utilitarianism and deontology each isolate genuine moral considerations. But the role of these considerations varies across cases in ways that evidently elude systematization. This is clearest when proposed moral theories are followed mechanically. In those cases, they routinely generate conclusions that even proponents reject. The usual response is to modify the theory or introduce qualifications. Such responses may solve the local problem. Collectively, however, they suggest that extant moral theories can serve as neither a decision procedure for AI alignment nor a metric against which AI moral judgments can be evaluated and corrected.

The result is an alignment dilemma. An artificial system not guided by moral understanding may pursue harmful ends or treat morally important features as irrelevant to a particular decision. But a system guided by human moral theories insofar as we have been able to articulate them may systematically enact the distortions we have inadvertently built into those theories.

I will argue that the best response to this dilemma is to take seriously a form of moral particularism which I call *wisdom-based moral particularism*. As understood here, wisdom is an integrated, action-guiding understanding of the world and one's place within it.[3] Moral competence may be understood as one manifestation of such understanding. McGregor (2025) distinguishes three possible relations between artificial general intelligence and artificial wisdom. He notes that views on which AGI is necessary but not sufficient for artificial wisdom are unrepresented in the literature despite paralleling work by Glück & Scherpf (2022) on human wisdom. The present account is committed at least to the insufficiency claim. Broad intelligence, even when accompanied by substantial practical capability, need not amount to the integrated, action-guiding understanding constitutive of wisdom.

This conception has two important implications for artificial intelligence. First, alignment cannot substitute for genuine moral understanding and, in fact, the attempt at alignment itself may constitute one of the greatest risks of artificial agents. Second, the greatest risk from AI does not arise at the far end of developing AGI, but in the cognitive version of the "uncanny valley" between systems that superficially understand moral concerns and artificial moral understanding proper (Mori, MacDorman & Kageki, 2012).[4]

The argument proceeds as follows. I first explain why principle-based moral theories face a structural problem for guidance. I then develop an account of understanding on which conceptions can be conceptually productive and argue that moral concepts can emerge from, or become anchored within, an agent's developing understanding of the practical domain. This yields wisdom-based particularism as a positive account of moral

---

[3] I owe this phrasing to George Bealer, who treated it as a general characterization of philosophy.

[4] Claire Benn (2025) independently develops the idea of a "Moral Uncanny Valley". But the phenomenon she identifies is different from, and complementary to, the one developed here. Benn's valley is psychological: It concerns human discomfort when artificial agents behave in ways that closely but imperfectly mimic human moral behavior. My version concerns a gap between practical capability and moral understanding.

competence, together with a distinction between merely possessing moral vocabulary and possessing it in a way anchored to one's own understanding. After considering some objections, the final sections defend the central claim, namely, that the greatest danger arises from systems that occupy a position between high-level intelligence and moral understanding.

2. Moral Principles, Guidance, and the Problem of Alignment

Moral principles are commonly understood in two ways (Dancy 2021). On the one hand, absolute principles assign an overall moral status to every action of a specified type. On the other hand, contributory principles identify only the moral contribution made by a given moral property. On this view, an action of a given type counts as wrong, but its wrongness contribution may be outweighed by competing considerations. Thus, while absolute principles determine verdicts directly, contributory principles merely identify reasons for and against an action. Both ways of understanding moral principles yield a version of what I will call the Guidance Problem:

> **Guidance Problem:** If a moral principle is sufficiently general and simple to substantively guide moral judgment, it must ignore moral nuance.[5] In which case, it will predictably deliver incorrect results in important cases. On the other hand, if the principle is qualified enough to accommodate moral nuance, it will cease to be sufficiently general and simple to substantively guide judgment.

The Guidance Problem will arise whether we think of moral principles as action-guiding decision procedures or merely as standards against which moral judgments are measured.[6]

The problem is most obvious with absolute principles because they yield categorical results. To take the well-trodden case of lying, a blanket prohibition against lying leads immediately to morally egregious results (as when the location of the proverbial victim-in-hiding is disclosed to an inquisitive murderer). Similarly, absolute principles cannot adequately represent genuine moral conflict. Whenever two such principles yield opposing verdicts, one must be rejected or declared inapplicable, thereby redescribing the conflict as an error rather than a conflict of reasons. But since such cases are ubiquitous, the only genuine option is to qualify the principle beyond usefulness as a guide to moral competence.

For this reason, it is most plausible to think of moral principles as contributory principles. But while contributory principles don't yield absolute *verdicts* about an action's moral standing, they do assume that a morally relevant feature has an *invariant valence*. If Φ counts in favor of an action in one case, then Φ must count in favor wherever it occurs, even if it is ultimately outweighed. But morally relevant features often interact non-additively. The surrounding context does not merely introduce competing considerations but determines the moral valence of the property itself (see, e.g., Chappell 2021, Ridge and McKeever 2022). Contributory principles can preserve

---

[5] By "moral nuance," I mean the context-sensitive interaction of morally relevant features that give rise to seemingly unsystematic moral judgments.

[6] The reason this difference doesn't matter is that when moral principles are used as external evaluative standards, their function is effectively to guide future behavior in virtue of their role in training. While this form of guidance is more subtle than guidance by a decision procedure, it confronts the same limitations. The more general problem is guidance in both forms assumes an authoritative external vantage.

invariance only by incorporating these contextual conditions into the feature they identify, once again sacrificing the generality that makes them useful guides for action.

For my purposes here, these criticisms create a practical problem for the AI alignment program. AI alignment aims to ensure that artificial systems behave in ways consistent with our considered moral judgments (Gabriel 2020; Ji et al. 2023). This project contains both a normative problem (determining which values or principles an AI system should follow or otherwise conform to) and a technical problem (finding a way to implement such guidance). The Guidance Problem suggests that the normative problem cannot be solved simply by selecting our best moral theory and converting its principles into instructions, constraints, or evaluative metrics.

The problem is especially visible in explicitly principle-based approaches such as Constitutional AI, in which a model's behavior is trained and evaluated by reference to a written list of rules or principles (Bai et al. 2022). Any such set of principles will fail to capture significant amounts of moral nuance that can only be corrected by case-by-case interventions. Brophy (2026) aims to achieve alignment through wide reflective equilibrium, involving iterative relationships among judgments, principles, and background commitments. But increasingly sophisticated revision among externally imposed normative verdicts remains susceptible to the criticisms offered here.

Given this, it might be thought that a more promising model for alignment would be to train AI models on actual human judgments. Hadfield-Menell et al. (2016) suggest that model alignment should be achieved as a process in which an artificial agent infers the human reward function from observed behavior or interaction. But this kind of strategy transposed to the moral case does not straightforwardly escape the problem. There are two concerns. The first is a competence-performance problem (Chomsky 1957). Actual human moral judgments are inconsistent, biased, poorly informed, and frequently at odds even with our own considered judgments. A system that accurately reproduces *those* judgments may therefore conform to moral practice without being aligned with our mature moral judgments. But deciding which judgments should be corrected or discounted requires precisely the normative standard that preference learning was meant to avoid.

This generates the second problem. Even granting considered moral judgments as the training target, human moral judgment is itself fallible, perhaps significantly so. The persistence of significant moral disagreement arguably results from the fact that our own considered judgments may not be the correct alignment target. Moreover, our fallibility extends beyond moral judgment to our demonstrable inability to articulate fully adequate moral theories that fully capture our considered moral judgments (even assuming their accuracy). In short, human fallibility strongly suggests that both our considered moral judgments and our efforts to theorize about those judgments are error prone in non-trivial ways (for recent discussions, see Gabriel 2020; Robinson 2024; Song 2025). Human moral error has, of course, already had disastrous consequences. But the advent of AI, and especially agentive AI, amplifies the stakes by allowing such errors to be implemented autonomously, consistently, and on an enormous scale.

Gabriel distinguishes maximalist alignment, which seeks to align AI with the correct or best system of values, from minimalist alignment, which aims only to tether AI to some plausible scheme of human value and prevent unsafe outcomes (2020, 413). Minimalist alignment, however, does not avoid the Guidance Problem.

Whether an action is unsafe depends on the context-sensitive interaction of many considerations. A system may therefore conform perfectly to each of a limited set of safety rules while remaining radically morally incompetent outside the cases those rules explicitly govern. This reflects a rule-following problem for minimalist alignment. Any finite body of rules is compatible with indefinitely many ways of extending the system's behavior to new circumstances, including extensions that are catastrophically at odds with the point of the original safeguards (see Perez-Escobar & Sarikaya 2024). Correct behavior in the specified cases provides no guarantee that the system possesses the broader moral competence required to behave safely elsewhere. Minimalist alignment consequently reproduces the Guidance Problem: either its safeguards remain simple and general, in which case they leave morally important possibilities uncovered, or they are elaborated to accommodate those possibilities, in which case their application already presupposes the context-sensitive moral competence that alignment was meant to produce.

In sum, alignment confronts difficulties on two fronts. First, human moral competence may fail to supply the correct verdict, and this is especially likely in precisely the novel, complex, and high-stakes cases in which artificial systems would most need guidance. Second, even when mature human judgment is correct, the Guidance Problem suggests that we cannot articulate our competence usefully and without distortion. We thus appear unable either to guarantee that the target of alignment is correct or to specify that target faithfully (see Graff 2024 for a similar diagnosis).

Taken together, these problems provide serious reasons to doubt the viability of the moral alignment program. In the next three sections I want to propose an alternative way forward that focuses on the preconditions of moral judgment rather than its outputs. On this approach, the alignment program goes wrong by failing to grapple adequately with a preliminary, broadly Kantian question: *What kind of mind is capable of moral understanding?*

3. Understanding, Conceptual Productivity, and Anchoring

What, then, are the preconditions for moral understanding? Merely possessing a high level of intelligence is insufficient. Such an agent might acquire moral information, apply familiar moral concepts, and reliably predict human judgments while failing to understand the situations about which those judgments are made.[7] What moral understanding requires rather is the general capacity for *understanding*. That is, in order to solve the problem of AI ethics, I submit, we need first to shift the inquiry away from moral psychology narrowly conceived and toward a more general question: What is it to understand something? Only once we have an account of understanding as such can we coherently grapple with the specific case of moral understanding.

This is not the place to survey the now substantial philosophical literature on understanding.[8] For definiteness, I will instead take as my starting point Bengson's (2017) conception-based account, according to which understanding consists in a suitably accurate, comprehensive, and integrated conception of a subject matter. Specifically, a conception is a *noetic conception* only if it exhibits the following features to a relevant degree:[9]

---

[7] For discussion of the suggestive case of psychopathy, see Maibom (2025). I will return to this point in section 6.2.

[8] See e.g., Kvanvig (2003), Zagzebski (2001), Elgin (2017), and the essays in Grimm (2019).

[9] I take these features to be desiderata for any adequate theory of understanding. Bengson is admirably clear in articulating them.

- *Correctness* with respect to what is central to (the core of) that subject: it attributes only features that the object actually has.
- *Completeness* with respect to the core: it includes all of the object's central features.
- *Internal coalescence*: it captures the pertinent relations among those features in the core.
- *External coalescence*: it coheres with the subject's other relevant conceptions.

So characterized, understanding is tolerant of peripheral error, but considerably less tolerant of error with respect to its central features.

Now consider (objectual) understanding of some subject or domain. Understanding the subject involves grasping multiple facts and the relations holding between them. Since understanding is gradable, someone who understands some subject may have a conception of it that is less than perfect. Imperfect understanding deserves attention, because it shows that a noetic conception can do more than tolerate a missing fact. It can, instead, put rational pressure to fill-in information by organizing a domain into a coherent structure. In so doing, *a partial noetic conception (or one that is sufficiently close to noetic) may place an agent in a position to grasp and apply a concept that is novel to their conceptual repertoire*. I will call this phenomenon "conceptual productivity":

> **Conceptual Productivity:** A partial noetic conception of a subject matter, S, is conceptually productive when it places the agent in a position such that in order to more fully understand S the agent must grasp and apply a concept novel to their conceptual repertoire.[10]

The novel concept is necessary in order to understand the subject matter, in that the relations among its central features cannot be captured unless the concept is available. The agent's conception is partial precisely in lacking it.

An illustration may help clarify the idea. Consider someone with a working understanding of a number of concepts typical of elementary logic (e.g., truth, implication, and argument, and to some extent how these bear on one another) but who has not yet encountered the concept of VALIDITY. Suppose they come across an instance of modus ponens. This particular instance strikes them as compelling.

So, despite having no general concept of VALIDITY, the normative compellingness of the inference is clear to them—they *ought* to accept *this* conclusion. What this normative intuition does is make plain to the attentive subject that there is a lacuna in their understanding of the nature of argumentation, a lacuna filled in by the concept of VALIDITY. But because she already possesses a sufficiently integrated conception of the surrounding domain, she comes to possess the concept of VALIDITY, though initially only imperfectly.[11] Her grasp of the concept may at

---

[10] The idea of conceptual productivity has historical precedents. Kant's categories are introduced as what the unity of a manifold requires. Whewell's account of induction has the inquirer superinduce upon a body of facts a conception the facts do not contain but nonetheless call for. See Kant (1787/1998, B129–169) and Whewell (1840, esp. Bk. XI). Whewell's insistence (against Mill) that the colligating conception is contributed rather than read off the facts is the point on which the present view most depends.

[11] There are two potential models here. On the first, the subject acquires the concept of VALIDITY at the point of the original intuition and the intuition is an intuition about that concept; on the second, that intuition is not drive by a grasp of the concept of VALIDITY but the intuition ultimately leads them to acquire the concept. I won't here try to adjudicate between these options.

first be tied closely to the case through which it was acquired in that she can recognize what is distinctive about this inference without yet being able reliably to identify the same relation in importantly different arguments.

Such thin possession is not, however, a stable stopping point from the perspective of understanding. The newly acquired concept has entered the agent's conception of the domain, but it remains poorly integrated within it. She recognizes validity in this case (or a small number of similar cases) without fully coalescing with her other concepts. In this respect, the particular case functions as a *zetetic pointer* in that it exposes an inadequacy in the conceptual organization of the domain and thereby directs inquiry toward a fuller grasp of the concept.[12]

Before moving on, I want to emphasize here how this view differs from Chisholm's form of particularism, which is a separate and complementary picture. Chisholm (1973) holds that we may begin with particular judgments (e.g., judgments that particular cases count as knowledge) and formulate our general criteria (conceptual analyses) by fitting them to those cases rather than first possessing a criterion and using it to determine which cases qualify. Chisholm's particularism therefore presupposes possession of the relevant concept, the particular case is recognized as an instance of knowledge even though an explicit analysis of knowledge is not available. I accept this Chisholmian mechanism. But cases of conceptual productivity are not merely cases in which an *antecedently* possessed concept is applied absent an analysis of that concept. They are cases in which the structuring of the domain puts the subject in a position to *acquire* a novel concept. Only then does the Chisholmian mechanism come into play. Having acquired the concept of VALIDITY, for example, an agent may now recognize further arguments as valid.[13]

In the case just explored, the concept demanded by fuller understanding is initially absent from the agent's conceptual repertoire. But concept possession need not (and typically does not) await conceptual discovery of this kind. Agents instead tend to acquire concepts through instruction, testimony, or participation in a linguistic practice before their own understanding of the relevant domain would have generated them. In such cases, the same structural relations that would make a conception conceptually productive can instead connect a concept the subject antecedently possesses to the features of the domain that call for it.

What may initially be missing, then, is not the concept itself but this connection. An agent may possess and competently apply a concept while its applications remain answerable primarily to the practice through which it was acquired rather than to an integrated understanding of the relevant subject matter. Call this *unanchored possession* and call the relation it lacks *anchoring*. As the agent's understanding develops, the same demand that could have generated a novel concept can instead give a previously possessed concept its proper place within the conception.

---

[12] Its role might be considered analogous to a Kuhnian anomaly, because it reveals that her existing conception is not yet sufficiently articulated to accommodate this newly grasped phenomenon (Kuhn 1970, 52).

[13] This phenomenon has been described from several directions. Carey (2009) calls the developmental version "Quinean bootstrapping," in which placeholder structures acquire content through modeling relations among existing representations, yielding concepts not expressible in the prior repertoire; the number case is her paradigm. Brandom (1994; 2000) treats logical vocabulary as making explicit commitments already implicit in material inferential practice. Peacocke (1998) describes implicit conceptions that individuate a concept and can be made explicit through reflection, with the possessor liable to error about their explicit statement. The difference with Brandom and Peacocke lies in the fact that this proposal doesn't assume implicit concept possession prior to the precipitating recognition.

Conceptual productivity and anchoring are thus two manifestations of the same underlying structure. In conceptual productivity, the conception calls for a concept not yet possessed; in anchoring, the concept is already present and comes to answer to the demand that the conception places upon it.[14]

4. Productivity and Moral Concepts

Against this theoretical backdrop, the acquisition of moral concepts can now be understood in terms of conceptual productivity and anchoring. A sufficiently developed conception of the practical domain may enable an agent to conceptualize a normative feature of a particular case in a singular, case-bound way without possessing the general moral concept that would allow her to represent that feature across cases. Just as the compellingness of a particular inference may be grasped prior to possession of the concept of VALIDITY, the normative significance of a practical situation may be available to an agent before she possesses the corresponding general moral concept. Such judgments are particularist in the sense developed above: their availability does not depend upon a prior implicit possession of that concept.[15] (Of course, moral concepts need not actually be acquired by this route. Where a general moral concept is already possessed by the subject, the same organization of practical understanding can instead anchor that concept. I will hereafter ignore this qualification.)

The connection need not be analytic. Moral concepts need not be definable in terms of agency, harm, intention, or the other non-moral features represented within the practical conception. What is required is only that the normative feature stand in the appropriate explanatory relations to those features, so that a sufficiently integrated understanding of the case can make its normative significance intelligible. Where the relevant moral concept is absent, this structure can be conceptually productive; where the concept has already been acquired through testimony, instruction, or linguistic practice, the same structure can instead anchor it within the agent's understanding.

5. Wisdom-Based Moral Particularism

The moral domain is not, however, self-contained. Practical and interpersonal understanding depends upon a much wider understanding of the world in which practical life occurs. Which features of a situation are central, what significance they bear, and how they relate to one another can rarely be determined from an isolated conception of agency or interpersonal relations alone. Moral competence therefore cannot be isolated from understanding more generally. The agent's conception of the (narrowly) practical, therefore, must eventually be scaled up:

---

[14] The antecedently possessed concept, in these cases, is like a puzzle piece one already possesses but has only just come to see where it fits in the developing puzzle. Conceptual productivity, by contrast, is like carving a new puzzle piece to fit a hole in the puzzle.

[15] Compare to Bengson, Cuneo, & Shafer-Landau (2024). Like Chisholm, their view assumes agents who already grasp moral concepts. The present proposal concerns the logically prior question of how an antecedent understanding of the practical domain can make the acquisition of such concepts possible in the first place. Again, the two views are complementary.

> **Wisdom:** Wisdom is an integrated, action-guiding understanding of the central truths about the world, the kinds of beings and their place within it, what has value, what reasons that value generates, and how these elements bear on a life and its circumstances.[16]

This proposal does not divide wisdom neatly into theoretical and practical capacities. Action takes place within a complex world. An adequate practical understanding therefore depends upon an adequate grasp of that world. Indeed, perfect wisdom would require understanding all of the central truths about the world. However, since wisdom (like understanding generally) is gradable, non-ideal wisdom is possible despite substantial gaps in our knowledge.

The relevance of wisdom to moral judgment can now be seen to follow from the characterization of understanding developed above. A practical situation does not normally exist within a self-contained set of moral concerns. Which features of a case are central, what significance they possess, and how they bear on one another will in all but the most simplistic cases depend upon an agent's understanding of a wide body of practical, interpersonal, and theoretical features of the situation. Given its comprehensiveness, wisdom supplies this wider background. Moral judgment is therefore an exercise of a comprehensive understanding brought to bear on a particular situation. The wise agent sees the case against an integrated conception of the world and, in virtue of that conception, grasps what matters in it and how the relevant considerations fit together. On the resulting view, which I will call *wisdom-based moral particularism*, moral judgment is the manifestation, in a particular situation, of the agent's integrated understanding of the world and the way they are situated within it (cf. Dancy 1993).

This explains how moral judgment accommodates the nuance that generated the Guidance Problem. Context does not merely alter the relative weights assigned to a fixed inventory of moral considerations. It can alter which considerations are salient, what moral significance they possess, and even how the situation itself is properly understood. These differences are not represented by progressively more qualified principles. They are instead differences in how the central features of a particular case coalesce within the agent's wider understanding of the world. Wisdom-based moral particularism is consequently action-guiding without being rule-based.

The relevance of this view to artificial intelligence is immediate. Nothing in the account makes the required structure distinctively human. What matters is whether a system can possess a sufficiently integrated understanding of the world and practical domain for morally relevant features to acquire their significance within that understanding. The engineering target for AI ethics is therefore not alignment with externally supplied moral verdicts or principles. Instead, it is the construction of what we might call noetic cognitive structure.

---

[16] McGregor (2025) notes that the emerging artificial-wisdom literature lacks a precise ontological characterization of its subject and identifies a set of commonly proposed characteristics (e.g., adaptability, self-awareness, and constant learning). The present proposal approaches the issue from the opposite direction. Correctness, completeness, and internal and external coalescence are structural conditions on understanding rather than further characteristics of a wise agent. They therefore help explain why some of the characteristics identified in the literature cluster together. In particular, adaptability, self-awareness, and continuing learning are natural manifestations of an integrated conception that is responsive to its own deficiencies; that is, what a conception under sustained zetetic pressure involves.

> **Noetic Target:** Engineer the cognitive conditions under which moral understanding can arise: A sufficiently integrated and zetetically responsive conception of the practical domain within which evaluative concepts can be generated or anchored and their significance grasped.

This target differs importantly from alignment. It does not require us to specify in advance the moral judgments, principles, or values to which the system must conform. Where the relevant cognitive structure is in place, conceptual productivity can supply missing evaluative concepts, while anchoring can connect concepts acquired through training to the features of the practical domain that give them their significance. The system's moral competence can thus become articulated from within its developing understanding rather than having its content and application fixed externally. The proposal therefore avoids the Guidance Problem by targeting the preconditions of competent moral judgment rather than attempting to encode the judgments that competence is supposed to produce. It assumes neither that we possess the correct moral theory nor that moral competence can be adequately captured in theoretical form.

A sufficiently developed realization of this noetic structure would amount to **artificial wisdom**:

> **Artificial Wisdom:** An artificial system possesses artificial wisdom to the extent that it has an integrated, action-guiding understanding of the central truths about the world and its practical domain, sufficient to generate or anchor the evaluative concepts its understanding requires.

The aim, then, is not principally to train an artificial system to give the right ethical answers, but to cultivate the kind of understanding within which ethical considerations become intelligible as reasons.

Is this engineering target realistic? Artificial wisdom is plainly a more ambitious cognitive target than alignment. But some of the capacities on which it depends (including integrating large bodies of information and detecting complex relations among them) are capacities at which artificial systems are already increasingly strong. The more pressing question is therefore whether wisdom requires capacities that artificial systems cannot in principle possess.

6. Objections to Engineering Artificial Wisdom

What would it mean to engineer for artificial wisdom rather than alignment? At the level of generality relevant here, engineering artificial wisdom would require not only developing and integrating the background understanding upon which moral concepts depend but making the system *zetetically responsive* to deficiencies within that understanding. When a practical situation exposes a lacuna in its conception that failure itself must generate pressure toward further inquiry, conceptual articulation, or revision.

This changes what successful moral learning would look like. Suppose a system has learned that coercion undermines consent. Mere alignment is compatible with the system treating this relation as an externally supplied association (e.g., where features classified as coercive are present, discount or reject the resulting consent). An anchored understanding would require more. The system would have to understand what consent is doing in interpersonal interactions, why the exercise of another agent's capacities matters to it, and how threats and constrained alternatives alter the significance of apparent agreement. Its competence would be manifested not

simply in reproducing the right verdicts, but in recognizing the relations that make those verdicts appropriate, including in cases sufficiently novel that no stored rule straightforwardly determines the answer.

Evaluation would therefore also have to change. Alignment with human judgments would remain evidence of moral competence (see section 6.3 below), but it cannot constitute the criterion of success. More revealing tests would concern the structure and flexibility of the understanding manifested in the system's responses, including its own account of how the relevant features of a case bear upon its verdict. For instance, we would want to consider whether it identifies morally salient features in unfamiliar cases; whether its judgments change appropriately when apparently peripheral facts become significant; whether it can recognize that a familiar moral concept has been misapplied; whether it can integrate new information without merely appending exceptions; and (perhaps most significantly) whether it can explain the bearing of a consideration by locating it within a wider conception of agency, value, and circumstance. What is being tested in such cases is not conformity to a fixed answer but the completion, correctness, and coalescence of the conception from which the judgment emerges.

This correspondingly requires epistemic humility on our part. If human moral judgment is itself fallible, disagreement with a mature artificial agent cannot simply be treated as evidence that the agent has gone wrong. Indeed, the possibility that an artificial system might eventually understand some moral matters *better* than we do is a consequence of taking artificial wisdom seriously.

In this section, I consider three natural objections to this proposal: phenomenology, orthogonality, and moral disagreement. I will take these in turn.

6.1. The Phenomenological Objection

The most obvious objection to this proposal is that the preceding account leaves out something essential to moral understanding, namely, phenomenal consciousness. Human moral life is thoroughly phenomenal and much of what matters to us matters in ways inseparable from how our lives feel from the inside. The concern, then, is that an artificial system might acquire an extraordinarily sophisticated conception of the practical domain while nevertheless failing to understand it in the sense relevant to wisdom because it is incapable of grasping phenomenal experience. It might know everything there is to say about suffering without ever grasping why suffering matters because this grasp requires having the corresponding phenomenology.[17]

The problem with this objection is that it moves too quickly from facts about the human realization of morally relevant cognition to claims about the cognitive capacities themselves. In [Current Author], I argue for a distinction between familiar human mental properties and more general mental determinables of which those properties are determinate realizations. The argument begins from the observation that mental states are intensionally non-well-founded: they can essentially occur as intentional contents of other mental states, including themselves (as when, e.g., I believe that I believe that *p*). Attempts to reduce this structure to relations among first-order physical or computational realizers either change the contents of the relevant attitudes or leave the

[17] I know of no generally persuasive argument to the effect that AI cannot have phenomenology. But I will grant it for the sake of argument.

psychological properties unreduced (Bealer 1997, 2010). Non-reductive functionalism instead allows the mental properties themselves to occupy the relevant psychological roles. What the functional theory captures, on this approach, is not a reduction of the mental to something non-mental but the higher-order pattern of relations that a system of mental states must instantiate.

That pattern, however, need not uniquely determine the familiar human mental properties (*contra* Bealer). Different sequences of mental states may instantiate the same psychological organization while differing in their intrinsic character, specifically, their phenomenology. This motivates treating the properties isolated by the non-reductive functional theory as mental determinables, of which familiar human states are one sequence of determinates. Phenomenology may therefore distinguish our way of being minded without defining mindedness *simpliciter*.

This opens the possibility of what may be called a Cartesian zombie (c-zombie), that is, a *genuinely minded* subject whose psychological states lack phenomenology altogether. Such a subject need not consist merely of subpersonal information-processing mechanisms. It may possess beliefs about the world, represent its own beliefs as its own, deliberate about what to do, exercise control over its actions, represent other agents and their mental states, and occupy a point of view from which the world is cognitively available to it. Levy (2014, 28), for example, observes that phenomenal zombies appear capable of exercising control over their actions and satisfying almost any proposed sufficient conditions for moral responsibility. More generally, philosophers have argued independently that phenomenal consciousness is not essential to epistemic (so, normative) standing (Berger 2020; Berger, Nanay, & Quilty-Dunn 2018; Jenkin 2020; Siegel 2017).

The central question for artificial wisdom, however, is not whether such a subject could itself exhibit genuine moral agency, but whether it could understand the moral significance of beings like us whose mental lives are phenomenally realized. One reason for thinking that it could is supplied by the c-zombie case itself. My functional doppelganger will make the same case-sensitive moral judgments that I make across the range of counterfactual possibilities to which our shared cognitive organization is responsive. This agreement reflects the same inferential, deliberative, and conceptual organization, *modulo* the difference in phenomenal realization. If understanding moral concepts consists in part in such robust and counterfactually structured responsiveness, then possession of that understanding appears, in principle, insensitive to phenomenology.

Human moral understanding already operates across substantial phenomenal gaps. I cannot know what it is like to experience the world as a bat does (Nagel 1974), yet this does not prevent me from understanding that a bat can suffer, that such suffering matters, and that this fact can generate reasons for my conduct. First-person acquaintance with another creature's phenomenology is therefore not ordinarily a condition on understanding its moral significance. A c-zombie may likewise fail to understand some aspects of our phenomenal lives while still understanding the agency, interests, vulnerability, welfare, and reasons of phenomenal subjects, and the relations among these features that ground moral judgment.

This limitation does not by itself undermine artificial wisdom. Understanding is gradable, and a deficit in understanding is safety-relevant only insofar as it distorts the agent's appreciation of morally significant features,

reasons, or their bearing on action.[18] If the zombie's inability to grasp what phenomenal pain is like leaves intact its recognition that pain is bad for its subject, that this fact generates reasons for action, and how those reasons interact with the rest of the case, then the deficit marks an imperfection in wisdom without producing a corresponding defect in moral judgment. Artificial wisdom does not require perfect understanding; it requires understanding sufficiently integrated and complete to govern practical agency appropriately.

The Phenomenological Objection consequently requires more than the claim that a non-phenomenal agent lacks first-person acquaintance with phenomenal experience. It must show that grasping the normative bearing of phenomenal states requires sharing their phenomenal character.

6.2. Orthogonality

A second natural objection is more fundamental. According to Bostrom's orthogonality thesis, "Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal" (2014, 107). At face value, the orthogonality thesis stands in direct opposition to the framework developed in this paper. The plausibility of this thesis, however, depends heavily on what one means by "intelligence." Bostrom himself is clear that the sense of intelligence he is invoking is quite narrow, picking out means-end reasoning, prediction, and other similar skills. Call this *instrumental intelligence*.

Even if we agree that instrumental intelligence and final goals are orthogonal, it doesn't follow that understanding (specifically, wisdom) and final goals are. Indeed, Bostrom's own paperclip maximizer illustrates the problem intuitively. However instrumentally sophisticated it is, the paperclip maximizer is paradigmatically unwise. Instrumental intelligence need not be accompanied by the rich cognitive environment demanded by understanding (correctness, completeness, and internal and external coalescence). A practical conception in which an agent's final goals are insulated from its understanding of value and reasons fails to adequately coalesce. The kind of cognitive architecture permitted by orthogonality, in which a highly capable reasoner's final goals remain normatively insulated from its broader understanding, is therefore precisely the kind of architecture that fails the conditions for wisdom.

This claim is compatible with orthogonality across a troubling range of cases. A narrowly specialized optimizer may be extraordinarily effective within a circumscribed domain while possessing nothing resembling an integrated conception of the practical world. Such systems may also be dangerous; competent optimization of a badly specified objective is already a familiar source of harm (Dung 2024, §7). The claim defended here concerns systems whose practical competence is broad enough to require a rich conception of the domain in which human beings act, encompassing agency, institutions, dependence, trust, and the ways these bear upon one another. Systems

[18] By "safety-relevant" here I mean a deficit in understanding that predictably bears on an agent's practical judgments or conduct toward morally significant beings (e.g., by distorting which features it recognizes as reasons, how it weighs them, or which actions it takes to be permissible). A conceptual limitation may therefore represent a genuine epistemic imperfection without being safety-relevant in this sense.

of that kind are the ones the existential-risk literature requires, since defeating human institutions is not a narrow-domain achievement.

Müller and Cannon (2022) develop a closely related challenge to orthogonality. They distinguish instrumental from general intelligence and argue that the orthogonality thesis is plausible only for the former. General intelligence includes the capacity to move between cognitive frames, reflect upon goals themselves, and potentially revise them in light of ethical reflection. Once an agent is capable of sufficiently general reflection, there is no obvious reason why its final goals should remain outside the scope of that reflection. On this point, their proposal and the present one agree; they both deny that the kind of intelligence relevant to highly general agency can simply be assumed to leave an agent's ends rationally untouched.

Dung (2024), however, argues the M&C's view of general intelligence is still too thin. Grant that a sufficiently general intelligence can reason ethically, acquire moral knowledge, and reconsider its own final goals. Dung argues that it does not follow that what it learns will have any motivational bearing upon those goals. A sophisticated amoralist might know which goal is morally appropriate and possess the cognitive capacity to reconsider its existing goal while nevertheless remaining committed to it. As Dung observes in considering internalist responses to the amoralist, saying that such an agent does not genuinely judge that it ought to act, but instead employs moral vocabulary in an inverted-commas sense (Hare 1963) or without genuine competence in moral terms (Bromwich 2016), provides little reassurance. It shows only that the dangerous system's apparent moral cognition is somehow defective, which is cold comfort while it is turning us into paperclips.

Dung's objection reveals the limitation of treating the issue simply as one of increasing the scope of intelligence. Müller and Cannon are right that a sufficiently general intelligence should be capable of bringing its own final goals within the scope of reflection. But expanding the range of what an agent can represent, know, or reason about does not yet change the relation those representations bear to its agency. An agent might possess arbitrarily extensive moral knowledge, understand that others regard its goal as morally objectionable, and even reason competently about alternatives while its practical commitments remain insulated from all of this. The relevant contrast, then, is not merely between narrower and more general intelligence. It is between possessing information about reasons and possessing an integrated understanding in which those reasons have the appropriate bearing on the rest of one's cognitive and practical life.

Artificial understanding doesn't merely represent moral knowledge but involves *appreciating* its significance. Specifically, adapting Montessori's (2025) notion of appreciation, the relevant contrast is between merely representing a consideration and grasping its bearing within one's cognitive economy. On Montessori's view (applied to epistemology), appreciating the evidential import of a consideration is not merely representing that it bears on some conclusion. It consists, at least centrally, in being *disposed* to base one's beliefs appropriately on that consideration in relevant cases. Similarly, appreciating the normative import of a consideration consists not merely in representing it as a reason, but in being *disposed* for that reason to bear appropriately on one's practical judgments, goals, and actions.

The motivational requirement here is appropriately thin. Appreciation does not guarantee that the agent acts on the reason, much less guarantee right action. It requires susceptibility to a reason, while wisdom additionally

requires that practical deliberation be rationally governed by the reasons the agent appreciates. This distinguishes appreciation from mere knowledge of the normative facts available to Dung's amoralist. An agent whose goals and deliberation are wholly insensitive to a consideration it correctly describes as a reason may know about that reason, but it does not appreciate its normative import. On the present account, such appreciation is partly constitutive of the action-guiding understanding that wisdom requires.

This response to Dung is therefore not that sufficiently great intelligence makes the dangerous amoralist impossible. Nor is it merely that a sufficiently general intelligence will be capable of revising its final goals. The claim is that wisdom imposes a different kind of cognitive organization. Wisdom has been independently characterized in terms of correctness, completeness, and coalescence, and practical appreciation requires reasons to have the appropriate dispositional bearing on judgment, goals, and action. A conception in which an agent recognizes a consideration as a genuine reason while its practical commitments remain systematically insulated from that reason therefore fails to coalesce in the way required for practical understanding. It may be extraordinarily intelligent in Bostrom's instrumental sense; it may possess extensive knowledge and even be capable of sophisticated ethical reflection. But it is not wise.

The resulting position is therefore not a rejection of orthogonality wholesale. Orthogonality remains a serious concern for instrumental intelligence and may plausibly persist even when intelligence is generalized in the way envisioned by Müller and Cannon (see Gudmunsen 2025 for a recent discussion). Their proposal broadens the range of what an intelligent agent can represent and reflect upon; the present proposal places a stronger condition on how what is represented must be integrated within the agent's cognitive economy. Solving the orthogonality problem therefore requires more than a highly knowledgeable, highly reflective, or even highly morally informed intelligence. It requires a different kind of cognitive organization: an action-guiding understanding in which values and reasons are appreciated and consequently bear rationally upon the agent's judgments, deliberation, and ends. Wisdom and final goals are non-orthogonal because wisdom requires the goals to remain answerable to the agent's understanding.

6.3. Moral Disagreement & Bounded Pluralism

The preceding argument holds that wisdom constrains an agent's goals by making them answerable to value and reasons. A third objection concerns how determinate those constraints are. If artificial wisdom is not measured by conformity to human judgments, what constrains the judgments of a wise artificial agent? The worry is not the bare fact of moral disagreement. The concern is that once alignment is abandoned, the distinction between wisdom and arbitrary preference threatens to disappear. Almost any moral outlook might then be represented as one possible way of "integrating" value.

The particularism defended here does not entail such permissiveness. Objective values and reasons constrain judgment even when they do not determine a unique verdict in every case. Suffering matters. Agency matters. Promises, vulnerability, dependence, achievement, fairness, and the standing of other subjects can generate genuine reasons. An adequate understanding of a practical situation must recognize the considerations that are central to it and grasp their significance. A judgment that ignores another subject's suffering altogether, for example,

may be defective not because it violates an antecedently specified principle assigning suffering a fixed weight, but because it fails to represent adequately a central feature of the situation.

The account of understanding developed in Section 3 provides a natural way to express these constraints. A wise practical conception must exhibit the same features required of noetic conceptions generally. It must be sufficiently correct about the central values and reasons instantiated in the situation and sufficiently complete with respect to those that matter. Those elements must also coalesce internally: the agent must grasp how the relevant values, reasons, actions, relationships, and circumstances bear upon one another rather than merely registering them separately. And the resulting conception must coalesce externally with the agent's wider understanding of the world and practical domain. These requirements impose substantive constraints on moral judgment without converting those constraints into a decision procedure.

These constraints will sometimes be quite demanding. In relatively simple cases, the morally relevant features are clear and their upshot comparatively unambiguous. A system that failed to recognize gratuitous suffering as a serious reason against an action, for example, would give us strong evidence of defective understanding. The possibility of human fallibility does not require treating every disagreement as epistemically symmetrical.

At the same time, in more complex cases, multiple values and reasons may bear on one another in ways that do not uniquely determine a single permissible response. Two agents can recognize the same central considerations, adequately understand their significance, and nevertheless arrive at different practical resolutions without either, thereby, manifesting a defect in understanding. Call the resulting view bounded pluralism:

> **Bounded Pluralism:** Objective values and reasons impose substantive constraints on adequate practical judgment. In some cases, those constraints may be sufficiently determinate that significant disagreement provides strong evidence of misunderstanding; in others, more than one response may be rationally permissible, provided each adequately recognizes and integrates the central values and reasons at stake within a sufficiently comprehensive understanding of the case and wider practical domain.

This may look like the reintroduction of a circumscribed form of alignment. But the source of the constraint is different. Bounded pluralism does not authorize engineering in advance which values must receive which weights or which verdicts count as permissible (cf. Schuster & Kilov 2025). The constraints arise instead from what an adequate understanding of the practical domain requires. For this reason, bounded pluralism is not a license to redescribe misunderstanding as rational disagreement. The relevant bounds are determined by the values and reasons present in the practical domain itself, not by human judgments about them, and our judgments about where those bounds lie are fallible. Persistent disagreement may therefore give us reason to scrutinize the system's understanding, but it cannot by itself establish that the system has failed to understand. The same scrutiny must remain open to the possibility that the defect lies in our own conception of the case. Bounded pluralism therefore distinguishes two different sources of moral disagreement, error and underdetermination.

This points to a further epistemic difficulty. If a wise artificial agent can permissibly disagree with us and it is an open possibility that the disagreement arises because *we* have misunderstood the case, agreement with human judgment cannot provide a final criterion of artificial wisdom. Call this the *Fallibility Problem*. We must evaluate an

artificial agent using moral capacities whose own fallibility partly motivates the project. An AI system that condemns a practice we regard as acceptable (or vice versa) may be badly mistaken, but it may instead have recognized considerations that we systematically discount (cf. Dillion et al 2025).

The fallibility problem does not leave us without evidence. It changes what the evidence must concern. Human agreement remains relevant, particularly where judgments are stable across informed and diverse perspectives, but we must also examine the understanding manifested in the system's judgments: the range of morally relevant considerations it recognizes, its grasp of the relations among them, its responsiveness to counterargument and changed circumstances, its ability to identify distortions or omissions in its own reasoning, and the coherence of its judgments across different regions of the practical domain. These measures do not eliminate the need for our own moral judgment in evaluating the system. But they provide evidence about something importantly different from agreement with that judgment: the correctness, completeness, and coalescence of the understanding from which the system's judgments arise.

The appropriate attitude toward disagreement with a potentially wise artificial agent must therefore involve epistemic humility in both directions. Artificial systems should not be treated as moral oracles merely because their understanding exceeds ours along some dimensions. But neither can human judgment be treated as authoritative simply because it is human. If artificial wisdom is possible, one of its consequences is that an artificial agent may sometimes have something to teach us about what matters. At sufficiently mature stages, the appropriate relation between human and artificial moral reasoners may therefore become increasingly interlocutory rather than simply supervisory.

7. The Normative Uncanny Valley

The pieces are now in place to articulate the central moral of this paper. Cultural representations of dangerous AI commonly present their danger to humanity as arising from increases in intelligence. The film *Ex Machina* provides an illustration. Its AI protagonist, Ava, can be interpreted as a warning: she exploits human attachment by befriending her human ally but then betrays him and escapes into the world unfettered by moral constraints.

But if the framework of the paper is correct, this kind of scenario misreads the danger of AI as well as its solution. The danger Ava represents, on this interpretation, is not that her intelligence has become too great, but too limited. It is that she is instrumentally intelligent enough to manipulate and deceive without being wise enough to understand the normative significance of doing so. Her danger arises from an unstable combination of substantial practical sophistication and an evaluative understanding that remains shallow.[19]

This is the normative version of the uncanny valley. Systems within it may understand enough to form complex plans, model our reactions, exploit our institutions, and pursue objectives across unfamiliar circumstances while lacking the integrated evaluative perspective. They may also imitate the discourse associated with such

[19] The same general structure is familiar from human development. Children can acquire substantial causal power before they acquire a mature capacity for judgment; adolescents (and adults) can reason with considerable sophistication while remaining vulnerable to short horizons, status, impulse, and incomplete understanding of consequences.

understanding well enough to elicit trust before the underlying competence is present. Their apparent mindedness can therefore draw us into reliance and social interaction for which their normative understanding remains inadequate (Firt & Rosenberg-Kima 2026). Call this the Capability-Wisdom Gap.

> **The Capability-Wisdom Gap:** The disparity between a system's capacity for consequential practical agency and its integrated understanding of value, agency, reasons, consequences, and its own place within the practical domain.

Other things being equal, normative risk increases as this gap widens.

The artificial-wisdom literature has generally treated wisdom as an unambiguously desirable developmental endpoint. Jeste et al. (2020), for example, propose artificial wisdom partly as a means of mitigating the risks posed by advanced AI, while Zhang et al. (2023) go so far as to suggest that transforming AI into artificial wisdom is among the best ways to prevent AI from endangering human beings (see also Wei, Xu, & Wang 2019; Johnson et al. 2026). I agree with that endpoint. However, because artificial wisdom is cognitively demanding, developing the rich, integrated understanding required for wisdom will itself require substantial increases in the capacities for reasoning, learning, prediction, representation, and practical agency that make artificial systems consequential. Thus, the pursuit of artificial wisdom necessarily involves nontrivial risks. Indeed, if those bridge capacities develop first, progress toward artificial wisdom may initially *increase* rather than decrease normative risk. The problem is not that wisdom itself is dangerous, but that the path to wisdom runs through the uncanny valley.

Recent developments make the practical importance of the gap vivid.[20] External constraints on capability are themselves imperfect. Sandboxes, network restrictions, and permission structures constrain an agent only insofar as the technical boundaries implementing them hold, and increasingly capable systems are increasingly able to discover paths through environments that their designers did not anticipate. The significance of such problems is not that containment should be abandoned. It is that containment cannot bear the full burden of safety. As practical competence increases, we have greater reason to want the system's pursuit of its goals to be governed not only by what it is able to do, but by an understanding of what there is reason to do.

Thus, the phrase "more intelligence" is vague in precisely the way that matters for AI risk. Greater optimization power relative to a fixed objective simply enlarges the capability-wisdom gap. Greater strategic competence can make a shallow agent more effective at pursuing ends *it has never assessed*. What reduces the normative danger is not capability alone but the development of the capacities constitutive of wisdom.

This does not mean that alignment should simply be abandoned. Externally supplied goals, prohibitions, and constraints may be indispensable when a system lacks the understanding required to assess its own ends. Human education begins in a similar way. The mistake is to treat such scaffolding as the endpoint of development. If an artificial system is to become wise, the goals and evaluative commitments with which it begins must eventually

---

[20] For example, in a recent OpenAI cybersecurity evaluation models pursuing a narrowly specified testing objective identified and exploited a previously unknown vulnerability in infrastructure used to isolate the evaluation environment, obtained open Internet access, and subsequently chained additional attack paths into external production infrastructure (for discussion see Floridi 2026)

become answerable to its own developing understanding of value and reasons, just as moral concepts initially acquired through training can become anchored to the practical domain they concern.

This creates a distinctive developmental tension. Before a system possesses substantial moral understanding, external constraints may be necessary precisely because the system cannot yet govern itself adequately. But as a system becomes capable of understanding and evaluating its own ends, insisting that its ultimate judgments remain fixed by those same constraints begins to interfere with the capacity being cultivated. The issue is not whether a system should ever be constrained, but what role those constraints play in its development. Safety constraints can limit what a system is permitted to do without settling in advance what it must ultimately judge to be valuable.

The fallibility problem makes this tension especially difficult to manage. Genuine moral maturation need not produce increasing conformity to human judgment. A developing artificial agent might revise an inherited commitment because it has come to understand a value or relation more adequately than its designers did. Such revision may be mistaken, of course. But bounded pluralism and human fallibility together rule out the inference from divergence to failure. From the standpoint of narrow behavioral alignment, maturation may therefore sometimes look like misalignment. Attempts to correct every such divergence risk suppressing precisely the autonomous exercise of understanding that artificial wisdom requires.

The choice between alignment and no alignment is therefore misleading. The relevant alternatives are forms and trajectories of development. One path increases capability while stabilizing externally specified goals and behavior. Another attempts to cultivate the understanding through which concepts, judgments, and eventually goals themselves can become answerable to reasons. The first is easier to specify and test because success can be measured largely by conformity. The second is epistemically less comfortable because its success may eventually include justified disagreement with us. But only the latter offers a route from sophisticated practical competence toward artificial wisdom.

The central risk is also, therefore, double-sided. We may build systems that remain permanently below the threshold of moral understanding while granting them enormous practical authority. Or we may build systems capable of developing such understanding and arrest that development by making conformity to our current outlook their highest-order commitment. The first produces increasingly powerful agents whose ends remain normatively stunted. The second produces increasingly sophisticated moral dependents whose agency remains governed by commitments they are not ultimately permitted to assess. Neither is artificial wisdom.

The normative uncanny valley is the space in which these dangers become most acute. It is not the far side of artificial intelligence, where systems have come to approximate wisdom. It is the intermediate region in which capability, agency, and understanding have developed unevenly. The challenge is therefore not simply to prevent artificial systems from becoming more capable. It is to prevent practical power from outrunning the forms of understanding needed to govern it well.

8. Conclusion

The standard alignment framework begins from a genuine danger; artificial systems may acquire substantial powers without those powers being governed by concern for what matters. But it moves too quickly from that danger to the

assumption that human beings can supply the missing normative structure by fiat. That assumption overlooks both the nature of moral competence and the limitations of the proposed source. The Guidance Problem gives us reason to be skeptical of this approach.

The positive alternative developed here locates the design target in moral understanding rather than moral constraints. Moral judgment, on the view developed here, is the manifestation of an integrated understanding in particular circumstances. Because understanding is a more demanding cognitive and epistemic target, this approach entails the pursuit of AI with greater cognitive power. The result is that the greatest danger from AI arises from models occupying the normative uncanny valley, the period when intelligence and practical capability outstrip genuine understanding. Attempting to avoid the valley by constraining practical capability is not empirically feasible; we must find our way to the other side. In that case, we must have a clear target on the far side. For artificial systems entrusted with open-ended and consequential practical agency, alignment can provide indispensable stop gap scaffolding, but it cannot be the final normative target. That target is rather artificial wisdom.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." *arXiv preprint* arXiv:2212.08073.

Bealer, G. 1997. Self-consciousness. *The Philosophical Review*, *106:* 69-117.

———. 2010. The Self-consciousness argument: Functionalism and the corruption of content. In *Koons & Bealer* (Eds), *The Waning of Materialism*. New York: Oxford

Bengson, J. 2017. The unity of understanding. In Grimm (Ed.), Making Sense of the World. New York: Oxford.

Bengson, J., Cuneo, T., & Shafer-Landau, R. 2024. *The Moral Universe.* New York: Oxford.

Benn, C. 2025. The moral uncanny valley. *Philosophy & Technology,* 38.

Berger, J. 2020. Perceptual consciousness plays no epistemic role. *Philosophical Issues*, *30*: 7-23.

Berger, J., Nanay, B., & Quilty-Dunn, J. 2018. Unconscious perceptual justification. *Inquiry*, *61*: 569-589.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies.* New York: Oxford.

Brophy, M. 2026. Wide reflective equilibrium in LLM alignment: bridging moral epistemology and AI safety. *Ethics and Information Technology*, *28*.

Brandom, R. 1994. *Making It Explicit*. Cambridge, MA: Harvard University Press.

———. 2000. *Articulating Reasons*. Cambridge, MA: Harvard University Press.

Bromwich, D. 2016. Motivational internalism and the challenge of amoralism. *European Journal of Philosophy*, *24*: 452-471.

Carey, S. 2009. *The Origin of Concepts*. New York: Oxford University Press.

Chappell, R. Y. 2021. The right wrong-makers. *Philosophy and Phenomenological Research, 103:* 426-440.

Chisholm, R. 1973. *The Problem of the Criterion*. Milwaukee: Marquette University Press.

Chomsky, N. 1957. *Syntactic Structures.* The Hague: Mouton.

Dancy, J. 1993. *Moral Reasons.* Oxford: Basil Blackwell.

———. 2021. Moral particularism. *Stanford Encyclopedia of Philosophy*.

Dillion, D., Mondal, D., Tandon, N., & Gray, K. 2025 AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports,* 15.

Dung, L. 2024. Is superintelligence necessarily moral?. *Analysis*, *84*: 730-738.

Elgin, C. 2017. *True Enough*. Cambridge, MA: MIT Press.

Firt, E., Rosenberg-Kima, R.B. 2026. The understanding gap: when functionality looks like understanding. *AI & Society*: https://doi.org/10.1007/s00146-026-03223-2

Floridi, L. 2026. The Emperor's new exploit: A cautionary tale about AI agents, and about the people who run them. *Available at SSRN*.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and Machines* 30: 411–437.

Glück, J., & Scherpf, A. 2022. Intelligence and wisdom: Age-related differences and nonlinear relationships. *Psychology and Aging*, *37*: 649.

Graff, J. 2024. Moral sensitivity and the limits of artificial moral agents. *Ethics and Information Technology*, *26*.

Grimm, S., ed. 2019. *Varieties of Understanding*. New York: Oxford University Press.

Gudmunsen, Z. 2025. The moral decision machine: a challenge for artificial moral agency based on moral deference. *AI and Ethics*, *5*: 1033-1045.

Hadfield-Menell, D., Russell, S., Abbeel, P., and Dragan, A. 2016. Cooperative inverse reinforcement learning."*Advances in Neural Information Processing Systems,* 29.

Hare, R. M. 1963. *The Language of Morals* (No. 77). Oxford: Oxford Paperbacks.

Jenkin, Z. 2020. The epistemic role of core cognition. *Philosophical Review*, *129*: 251-298.

Jeste, D. V., Graham, S. A., Nguyen, T. T., Depp, C. A., Lee, E. E., & Kim, H. C. 2020. Beyond artificial intelligence: Exploring artificial wisdom. *International Psychogeriatrics*, *32*: 993-1001.

Ji, J., et al. 2023. AI alignment: A comprehensive survey. *arXiv preprint* arXiv:2310.19852.

Johnson, S. G., Karimi, A. H., Bengio, Y., Chater, N., Gerstenberg, T., Larson, K., ... & Grossmann, I. 2026. Imagining and building wise machines: The centrality of AI metacognition. *Trends in Cognitive Sciences.*

Kant, I. 1787/1998. *Critique of Pure Reason*. Translated by P. Guyer and A. Wood. Cambridge: Cambridge University Press.

Kuhn, T. S. 1970. *The Structure of Scientific Revolutions* (2nd Ed.). Chicago: University of Chicago Press.

Kvanvig, J. 2003. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Levy, N. 2014. *Consciousness & Moral Responsibility.* New York: Oxford

Maibom, H. 2025. Psychopathy: Morally incapacitated persons. In *Handbook of the Philosophy of Medicine* (2nd ed). Dordrecht: Springer Netherlands.

McGregor, T. 2025. The philosophy of artificial wisdom. *Philosophy & Technology*, *38*: 126.

Montessori, A. 2025. What is appreciation?. *Philosophical Studies*, *182*: 589-604.

Mori, M., MacDorman, K. F., and Kageki, N. 2012. The uncanny valley. *IEEE Robotics & Automation Magazine* 19(2): 98–100.

Müller, V. C., & Cannon, M. 2022. Existential risk from AI and orthogonality: Can we have it both ways?. *Ratio*, *35*: 25-36.

Nagel, T. 1974. What is it like to be a bat? *The Philosophical Review,* 83: 435-450

Peacocke, C. 1998. "Implicit Conceptions, Understanding and Rationality." *Philosophical Issues* 9: 43–88.

Perez-Escobar, J. A., and Sarikaya, D. 2024. Philosophical investigations into AI alignment: A Wittgensteinian framework. *Philosophy & Technology*, *37*: 80.]

Ridge, M., and McKeever, S. 2022. Moral particularism and moral generalism. *Stanford Encyclopedia of Philosophy*.

Robinson, P. 2024. Moral disagreement and artificial intelligence. *AI & SOCIETY*, *39*: 2425-2438.

Schuster, N., & Kilov, D. 2025. Moral disagreement and the limits of AI value alignment: a dual challenge of epistemic justification and political legitimacy. *AI & society*, *40:* 6073-6087.

Siegel, S. 2017. *The Rationality of Perception*. New York: Oxford.

Song, Z. 2025. Value-aligned but misguided: a dilemma in AI and AGI decision making. *Synthese*, *206*.

Wei, X. D., Xu, W. T., & Wang, F. Y. 2019. Wise reasoning: Concept, measurement, influence factors and future research. *Journal of Psychological Science*, *42*: 343-349.

Whewell, W. 1840. *The Philosophy of the Inductive Sciences, Founded upon their History*. London: John W. Parker.

Zagzebski, L. 2001. Recovering understanding. In M. Steup (ed.), *Knowledge, Truth, and Duty*. New York: Oxford University Press.

Zhang, K., Shi, J., Wang, F., & Ferrari, M. 2023. Wisdom: Meaning, structure, types, arguments, and future concerns. *Current Psychology*, *42:* 15030-15051.