Meta-Reflective Capacities, Normative Commitments, and Responsible AI

Brendan Fleig-Goldstein

Department of Philosophy, Brown University, Providence, RI

brendan_fleig-goldstein@brown.edu

January 2026

Abstract

What capacities must an AI system possess to be held responsible for its actions? I argue that AI systems can be accountable agents when they possess sufficiently strong commitments to relevant norms (ethical, rational, or conventional). This paper articulates empirically determinable necessary and sufficient conditions for possessing such commitments. Specifically, I argue that what I term a meta-reflective capacity toward a goal is both necessary and sufficient. Meta-reflection is the capacity to maintain resource-optimal performance by appropriately changing one's cognitive strategy in response to changes in internal constraints. A commitment's strength can be characterized by the constraints under which a system fails to maintain resource-optimality. The path forward for a theory of AI responsibility requires articulating which qualitative internal system limitations are excusable, forgivable, or competence-undermining. Considerations of the mutability of constraints offer a partial way to delineate such classes. This framework connects philosophical theories of responsibility to cognitive processes and provides a path toward identifying and engineering normative commitments in biological and artificial systems.

1 Introduction

> "If it's possible for someone to lose their humanity, surely it must be possible for something that once had none to find it." Gene Wolfe

As efforts to build "agentic" and "autonomous" AI systems accelerate, so too does the need for a better understanding of machine agency and the conditions for moral responsibility

(Wallach and Allen, 2008). Currently, such artificial systems occupy a no-man's land of being powerfully intelligent and goal-directed, but not clearly competent enough to be fully self-governing participants in our normative practices (Cibralic and Mattingly, 2024). This no-man's land makes it hard to hold either the systems themselves or the companies that produce them responsible, and so the "responsibility gap" is worryingly widening (Cibralic and Mattingly, 2024; Matthias, 2004). What makes an agent normatively competent, and therefore a fitting subject for rights and responsibilities?

Human agents can claim competence and control in their decision-making and rightly accept praise or criticism for those decisions. This practice is highly valuable to human social life; people should want to be held accountable, as avoiding responsibility implies a lack of the control and freedom needed to enjoy many privileges associated with being human. Pleading incompetence to get out of a speeding ticket may save you the fine, but at the cost of your driver's license (Dennett and Caruso, 2021). Pleading insanity in court to avoid punishment means losing societal privileges. Shirking responsibility for reasoning mistakes means losing one's status as a competent and trustworthy epistemic agent.

The ability to be held accountable and enjoy related privileges requires a capacity to "adhere" to norms. A competent driver must adhere to the norms of the road; a moral agent to moral norms; and a serious epistemic agent to norms of rationality. But what is adherence?

Currently, AI engineers aim to build systems that behaviorally conform to established norms—such as traffic laws for self-driving cars or principles of fairness for law-enforcement algorithms. The reliance on general deep-learning training regimes to instill such conformance has recently been criticized by Millière (2025) as producing only "shallow alignment" between systems and norms. These alignment techniques reinforce surface-level behavioral dispositions—e.g., refusing certain prompts, echoing safety phrases—without endowing systems with the deliberative, normative reasoning capacities needed to resolve conflicts in novel cases. As a result, such systems remain vulnerable to prompt-injection attacks that exploit tensions between alignment norms (e.g., honesty and harmlessness).

If the goal is to engineer accountable AI systems, behavioral conformance alone is the wrong target. While some degree of conformance is of course required for normative competence, conformance per se is neither necessary nor sufficient for responsibility.

Behavioral conformance is not necessary—being held accountable for violating a norm requires that one is not conforming to it! A member of the "moral agents club" can make

errors, so long as they can be made to see them as such and respond appropriately (Dennett, 2015; Fischer and Ravizza, 1998; Strawson, 2008). As bounded agents, a perfect first-order disposition to conform to norms—never making any mistakes—is often infeasible and, in any case, unnecessary.

Nor is conformance sufficient. Accountability requires not only that an action stand in the right relation to a norm, but that it stand in the right relation to the agent. Philosophers have analyzed this agent–action relation in various ways: e.g., the action must arise from the agent's appropriate intentional states (Davidson, 1963); the action must be produced by a process that is responsive to reasons (Fischer and Ravizza, 1998); or the action must be grounded in the agent's normative commitments (Korsgaard, 2010).

Applying such philosophical accounts of moral responsibility to AI systems comes with difficulties. As will be argued in the next section, identifying internal mental states in AI systems is currently, and will likely remain, prohibitively difficult. Further, it is not clear how to identify the mechanisms from which actions are issued, nor how to determine whether such mechanisms count as appropriately sensitive to reasons.

The aim of this paper is not to offer a conceptual analysis of moral responsibility, but instead to articulate cognitive capacities that are necessary and sufficient for moral responsibility and can be tractably engineered and empirically identified in AI systems.

This paper adopts the view that responsibility and blame are best understood in terms of normative commitments and a system's capacity to hold them. The idea is that empirically determinable conditions can be spelled out that are necessary and sufficient for a system to possess normative commitments, without relying on speculative assumptions about mental states of AI systems. I argue that the relevant condition is possessing what I term a meta-reflective capacity toward an aim, which is the capacity to continue to perform optimally relative to constraints as those constraints change.

Following a discussion of normative commitments (§2), I introduce a resource-rationality framework (§3) to characterize the notion of meta-reflection (§4). The rest of the paper then supports my argument that this meta-reflective capacity is both necessary and sufficient for normative commitments, an argument I defend by analyzing cases from human, non-human animal, and artificial systems. I also discuss how establishing this claim helps focus the target for engineers who wish to create autonomous AI systems, and how it generally opens up new avenues for research into AI "minds."

	Complies with X	**Violates X**
Possesses a normative commitment to X	**Moral Agent** (Praiseworthy)	**Prima Facie Blameworthy Agent** (Morally Accountable)
Lacks a normative commitment to X	**Mere Conformance** (Not Praiseworthy)	**Incompetent System** (Not Morally Accountable)

Table 1: Normative Commitments and Agency: First Gloss

2 Normative Commitments

What is it to be committed to a particular norm? A normative commitment to X is more than a mere first-order disposition to X. Bilgrami puts the point like this:

> One must be prepared to have certain reactive attitudes, minimally to be self-critical or to be accepting of criticism from another, if one fails to live up to the commitment or if one lacks the disposition to do what it takes to live up to it; and one must be prepared to do better by way of trying to live up to it or cultivate the disposition to live up to it. (Bilgrami, 2008, p. 138)

I can acknowledge that I am not currently doing enough to combat climate change or help those in need, while still being committed to doing so. Similarly, I may exhibit various irrational behaviors while remaining committed to norms of rationality, as long as these irrationalities are genuine mistakes made in good faith, and I remain responsive to feedback.

Commitments are also not second-order dispositions (Bilgrami, 2008). A disposition to develop a disposition is not a commitment. You may not currently be disposed to compulsively scroll social media, but perhaps you are disposed to become disposed to scroll it. This does not mean you think you ought to. Potential addicts are not ipso facto committed to using. By similar reasoning, possessing higher-order dispositions of any degree is insufficient for possessing a commitment (Bilgrami, 2008).

If normative commitments are not merely dispositions or higher-order dispositions, what are they? Philosophical accounts of normative commitments are generally formulated in terms of reasons. For example, you have a normative commitment to X only if that commitment is a reason for you to believe or act in certain ways (Millar, 2004). If you're committed to avoiding animal products, that commitment becomes a reason to decline food that contains animal products. Commitments constitute reasons that a responsible agent must be responsive to.

Also formulated in terms of reasons is the concept of normative self-government: “our capacity to assess the potential grounds of our beliefs and actions, to ask whether they constitute good reasons, and to regulate our beliefs and actions accordingly" (Korsgaard, 2010, p. 6). This capacity presupposes normative commitments and the ability to reflect on and uphold them. While one arguably cannot have normative commitments without being normatively self-governing, the more relevant point here is that one cannot be normatively self-governing without normative commitments. A venerable tradition in philosophy has argued that agency and normative self-government are intertwined and even interdependent (Roskies, 2016; Silverstein, 2017).[1]

Thus, if we want “autonomous" (auto-nomos) AI systems–self-legislating systems–they will need to be endowed with the capacity for normative commitments.[2] But if normative commitments and normative self-government are understood in terms of reasons and are not reducible to behavioral dispositions, how can an AI engineer endow a system with a capacity for a normative commitment? The AI engineer designs causal systems; they do not work directly with reasons. Making causal systems sensitive to reasons has long been a goal of AI (Haugeland, 1981), but the preceding discussion illustrates the difficulty of this task, since merely inducing conformance to a first-order disposition is insufficient. It’s simply not clear, from a causal or mechanistic perspective, what it means for a system to be sensitive to reasons.

One solution to this challenge is to develop a framework for identifying and attributing mental representations to AI systems. For example, one could analyze normative commitments as a species of belief—namely, beliefs about what ought to be the case. Alternatively, one might hold that systems are responsible only for intentional actions (Strawson, 2008), where actions count as intentional only if they arise from appropriate beliefs and desires (Davidson, 1963). On either view, creating responsible AI would require endowing systems with genuine intentional states. This effectively replaces the problem of determining what makes an AI system responsible with the problem of determining the conditions for an AI system to possess belief and desire states (or perhaps weaker forms of representations).

[1]This idea goes back at least to Kant, who held that freedom, agency, and reason were all intertwined.

[2]Researchers such as Long et al. (2024) and their colleagues argue there is a “realistic, non-negligible possibility" that near-future AI systems will become subjects with their own interests who deserve moral consideration. They posit that “robust agency"—the capacity to pursue goals via belief- and desire-like states—is one of two primary routes to this status, even in the absence of accountability. Note that this paper is concerned with the more demanding sense of agency, i.e., identifying the necessary cognitive foundation for a system to be a fitting subject for praise and blame and therefore also certain privileges not granted to mere moral patients.

Probing beliefs and desires in AI systems would be enormously helpful for attributing responsibility and agency. For example, if an AI holds a belief but also desires to deceive—and thus asserts something contrary to that belief—it could be seen as intentionally lying. More generally, if normative commitments are just beliefs about what one ought to do, then being able to read and write such intentional states (e.g., a belief that one ought not to lie) could offer a path to identifying and inscribing commitments and empirically answering questions of responsibility. Normative commitments are rich mental states, so it would make sense to go "inside" AI systems and investigate their mentality, if any, to establish that they have such commitments.

Unfortunately, this approach simply trades one hard problem for another; identifying intentional content in AI is and will be, I claim, infeasible for the foreseeable future.

Consider Levinstein and Herrmann (2024)'s critique of attempts by Azaria and Mitchell (2023) and Burns et al. (2022) to infer belief states from internal model embeddings. They show that, *even if* LLMs possess beliefs, current probing methods are both conceptually and empirically inadequate. Probes fail to generalize across even basic transformations like negation, and there's reason to think that embeddings are too opaque, high-dimensional, and context-sensitive to reliably encode belief-like states. Probing seeks a stable, decodable signal of belief. Levinstein and Herrmann (2024) argue, however, that there's no reason to expect that such a context-independent feature corresponding to "truth" (thus indicating the LLM's endorsement of the proposition) exists in LLMs at all. This worry has been borne out by subsequent empirical work demonstrating that truthfulness-detecting probes may be locally effective but fail to generalize across tasks (Orgad et al., 2024). Linear decoding methods, or even more sophisticated probing techniques, are therefore unlikely to allow us to infer mental states in LLMs any time soon.

Or consider Hofweber et al. (2024)'s Minimal Assent Connection (MAC), which aims to infer mental states (namely, personal probabilities) directly from an LLM's next-token probabilities computed from its logits, rather than from internal embeddings. On this view, a model's credence in a proposition is the probability of outputting assent to a sentence expressing that proposition, after excluding irrelevant responses and normalizing. But as the authors themselves acknowledge, this only works if the model is assumed to be purely truth-seeking (and reporting) and lacks competing goals. In other words, as the authors state, a framework that identifies credences with next-token probabilities cannot account for what it means for a model to lie (Hofweber et al.,

2024). More broadly, when an LLM has goals other than reporting the truth—such as achieving its own objectives or sparing someone's feelings—the next-token probabilities reflect merely what it is likely to say, not what it believes. Such frameworks, therefore, presuppose a radically restricted form of agency. They are therefore inapplicable to capturing the kinds of intentional states required for responsibility in agents with complex, goal-directed behavior.

Indeed, identifying intentional content in *humans* by probing brain states is still highly controversial. Even setting aside complex debates over naturalistic theories of mental content, current philosophical analyses of mental content in cognitive neuroscience almost always provide accounts only of subdoxastic representations (i.e., concepts such as DOG, CAT, NUMEROSITY) (Neander, 2017; Shea, 2018). Frameworks for identifying full propositional content remain a promissory note (Neander, 2017; Shea, 2018). Subsequently distinguishing different attitudes towards a proposition to arrive at full intentional content is an even more distant prospect; while a neural decoder might eventually identify a representation of the proposition that the dog is on the mat, determining whether the subject believes, desires, or dreads this state of affairs remains far more challenging. Such controversy in the human case makes probing for content in the comparatively alien domain of AI all the more dubious. Thus, any strategy for recognizing or engineering AI agency that depends on identifying or inscribing intentional states within a model's internal architecture is and will likely remain infeasible for the foreseeable future.

One way out of this challenge is to adopt an ascriptionist approach to intentionality (e.g., Dennett, 1989). On this view, rather than looking inside a system to determine its representational content, it is sufficient that attributing beliefs and desires to the system grants us predictive and explanatory power over its behavior. This avoids the thorny issue of probing the internal states of AI systems. Ascriptionism is also appealing because it does not deny the possibility of identifying explicit internal representations; it simply holds that the status of any internal state as a representation is ultimately determined by its contribution to the predictable behavioral patterns that license the intentional stance in the first place. The identification of such internal vehicles is therefore a partly empirical question that need not be resolved before attributing intentional content to the system as a whole (Dennett, 1989).

Ascriptionism offers a programmatic framework rather than a specific answer to what makes an AI system responsible. What specific kinds of behavior should we look for to attribute beliefs or normative commitments? As discussed, mere first-order or even higher-order behavioral dispositions seem insufficient to capture normative commitments.

My approach strikes a middle ground. Like ascriptionists, I hold that criteria for responsibility can be articulated in behavioral terms. However, I argue for richer behavioral concepts than the simple dispositions of traditional behaviorism. To this end, I introduce the concept of meta-reflection, which identifies behavioral patterns that, I argue, are both necessary and sufficient for the cognitive capacities underlying normative commitments, and thus for agency and responsibility.

It may be that explicit internal intentional states are necessary for moral responsibility. I remain neutral on that question. But if so, my claim is that meta-reflection is sufficient to identify the presence of such states behaviorally, without first locating their internal vehicles—just as recalling a certain amount of information on a test reveals a minimal memory capacity without identifying its neural basis. Meta-reflection entails the relevant sufficiently sophisticated cognitive capacities and is therefore sufficient for grounding ascriptions of moral responsibility.

This project aims to provide an account of how to evaluate a system as normatively competent. In doing so, it complements but moves beyond recent work on responsible AI that focuses on minimal agency. For instance, Cibralic and Mattingly (2024) propose that simple goal-seeking behavior, regulated by a minimal internal representation, is sufficient to identify a system like a thermostat as a causally responsible agent. While their theory helps address the responsibility gap, it distinguishes this minimal agency from an account of moral competence; the appropriate response to a thermostat's malfunction is repair, not moral condemnation. My account attempts to specify the conditions for precisely this richer, normative competence.

I am not providing a conceptual analysis of normative commitments, and I remain neutral on a potential a posteriori identity claim between normative commitments and meta-reflection.[3] The claim is only that meta-reflection provides necessary and sufficient conditions for the presence of a normative commitment.

This proposal is analogous to Newell and Simon (1976)'s physical symbol system hypothesis or their search-and-heuristics hypothesis. Simon and Newell did not offer a conceptual analysis or an a posteriori identity claim about intelligence. Instead, they made an empirical claim about the general conditions for a causal system to be intelligent: a system is intelligent if it can search a problem space using heuristics and identify potential solutions. They were not proposing a specific model of how to build such a system, but by providing general constraints on a solution,

[3]As Bilgrami notes, even if there is an a posteriori identity claim between normative commitments and some natural property, this would not constitute a naturalistic reduction of normativity as, due to his Fregean-Moore pincer argument, it would still need to be present in its non-naturalistic form in the Fregean sense that enters into the identity claim.

they broke down the problem of designing “intelligence” into the more tractable engineering challenge of designing subsystems that can “search” and “test”. Similarly, by specifying meta-reflection as the relevant capacity, the task of building and evaluating responsible systems becomes one of investigating which first-order cognitive processes can give rise to this higher-level capacity (a question I turn to in §6.4).

3 Resource-rationality

This section introduces a general framework of resource-rationality to provide the conceptual scaffolding necessary to define the notion of meta-reflection developed in the next section.

The notion of rationality adopted here is a three-place relation between a cognitive strategy, goal, and environment: what is the optimal cognitive strategy relative to a specified goal within a particular environment? Since different strategies perform better or worse across environments, rationality is not a matter of mental processes conforming to abstract rules. This view, therefore, comports with substantive, as opposed to structural, theories of rationality, such as accuracy-first epistemology, which treat norms like coherence as a means to a further end, i.e., truth (Joyce, 2009).

The framework makes rationality empirically determinable by operationalizing each component. Cognitive strategies are understood as behavioral functions (Anderson, 1989). What observable “choices" did the system make during the task? What moves did they make in the board game? What credences did they report? Did these choices lead to better or worse outcomes relative to the goal? This behavioral focus minimizes speculation and reflects a commitment to the idea that intelligence is for action (Icard, 2014). If one prefers, cognitive strategies can be thought of as psychological processes coarse-grained into partitions of behavioral equivalence.

Task-environments, meanwhile, are stipulated by experimentalists evaluating a system (Icard, 2023). Consider a psychologist designing a laboratory task with an artificial environment and an explicit goal: perhaps a game where the aim is to maximize monetary reward, or a belief-updating task where the aim is to achieve maximal accuracy. Framing environments and goals in this way avoids speculation about “normal” or “evolutionary” environments and sidesteps the task of discerning a system’s intrinsic goals. Systems are evaluated instead on their performance against external benchmarks in the specific scenarios under study (I use "system" as a neutral term for any entity being evaluated, be it human, animal, or AI).

Resource-rationality builds on this picture of rationality by relativizing it to cognitive limitations, making it a four-place relation between cognitive strategy, goal, environment, and

the constraints that bound performance. It thereby asks what is the best that can be done given a particular set of constraints. On this setup, the pool of available strategies is restricted to those possible for systems subject to such constraints—e.g., particular memory bounds or processing speed limits. The resource-rational strategy is the one that performs best relative to the goal and environment from among those in the pool (Icard, 2023).

While one could allow only certain kinds of psychological facts to count as constraints, in this framework, any material fact about a system can count as a constraint. These can be understood broadly to include not only stable architectural or biological properties such as the conduction velocity down an axon or an artificial neural network's number of layers, but also dynamic, occurrent material states like glucose levels or the token sequence in an LLM's context window.

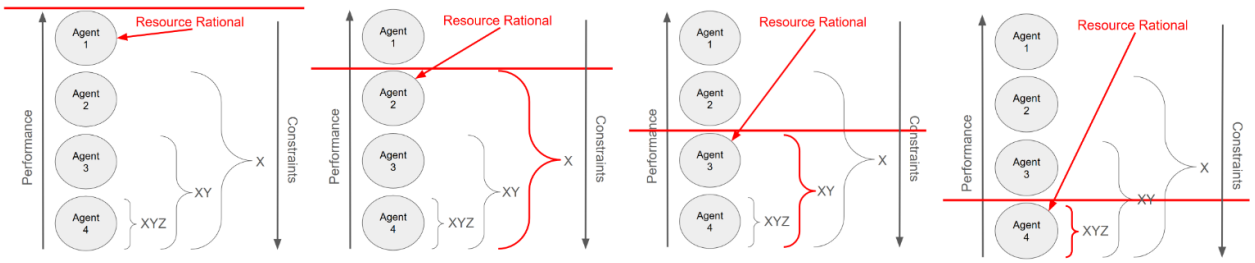

Figure 1: Resource-rationality or rationality relative to constraints. Relative to which set of qualitative constraints is a system (or agent) performing optimally? Each agent can be seen as performing resource-rationally, but relative to different constraints.

Given a broad conception of constraints, any cognitive strategy can, in principle, be understood as resource-rational, but to actually do so, one must empirically identify the specific material facts such that limiting the pool of available strategies relative to those constraints yields the observed behavior as the optimal one. Note that one cannot simply limit the pool of agents to those who behaviorally make a certain error; one must instead identify material facts about the system that give rise to such errors. This condition is what prevents meaningless analysis of all systems as resource-rational. §4.1 addresses how material facts are to be empirically identified.

In other words, this conception of resource-rationality is not focused on evaluating agents by labeling them as resource-rational or not, but instead on assessing *how constrained* systems are. The framework's goal is to bring into view all material facts preventing better performance. This step is prior to the subsequent step of deciding the normative status of constraints—i.e., which

constraints count as excusable, forgivable, competence-undermining, etc., as will be explained below.[4]

4 Meta-Reflection

Now it is possible to lay out the following notion:

> **Unbounded Meta-Reflective Capacity:** A property of a system such that intervening on any of its constraints results in the system remaining resource-rational.

Put simply, this is the capacity of a system to re-evaluate whether its approach still works as conditions change and switch strategies accordingly.

That is, if one modifies a system to be less constrained—e.g., by increasing its memory capacity—the system may need to adopt a new strategy to remain optimal relative to the now-expanded pool of possible cognitive strategies. While any system can be understood as resource-rational relative to a sufficiently large set of constraints, most systems do not exhibit meta-reflection for non-trivial goals: changing the system's actual constraints does not guarantee that it will revise its strategy. Unless it does so whenever the incumbent strategy ceases to be optimal, the system will fail to remain resource-rational.

When considering programs that can look up to five moves ahead, a particular heuristic for selecting chess moves may work extremely well relative to others, but not as well if it is possible to look fifty moves ahead. A linear bounded automaton (the finite tape version of a Turing Machine) might run a program that is optimal relative to its particular finite capacity. But unless it alters its program, adding more tape may result in it being suboptimal relative to this new amount of resource. In the other direction, further limiting the psychological resources of an agent may simply cause a breakdown in that agent's behavior (e.g., running off the tape), as opposed to a switch to a new, more appropriate strategy. In either case, the failure to modify its strategy is itself a product of an agent's psychological constitution and reflects a constraint. Thus, one must identify this further constraint and must further restrict the pool of agents by including this new constraint to bring the system's resource-rationality back into view. So again, all agents are resource-rational, but not all display meta-reflective capacity.

[4] I defend the particular conception of resource-rationality laid out in this section in several forthcoming papers; one motivation is an epistemological argument against drawing a principled line between types of psychological constraints; another is a methodological argument that this conception of resource-rationality provides a more effective strategy for studying human cognition.

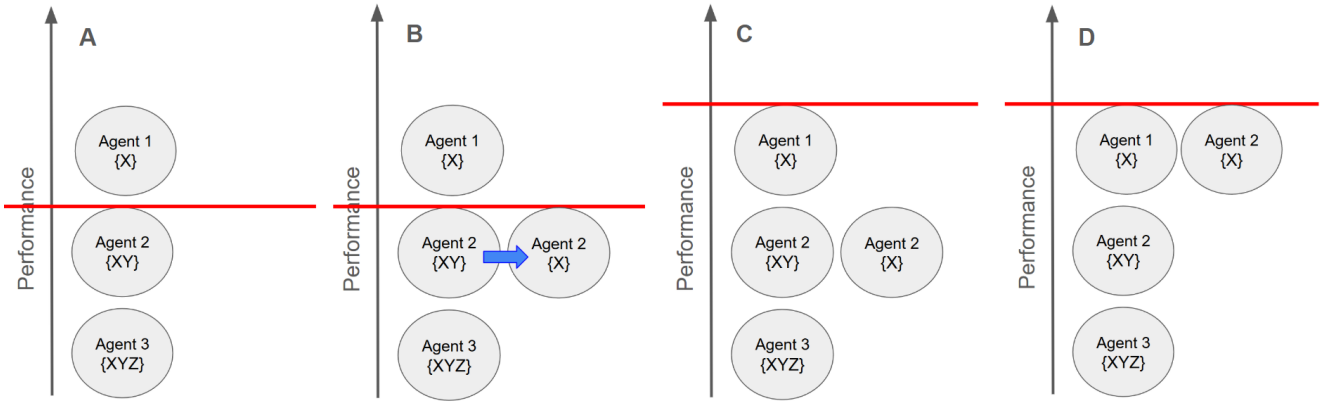

Figure 2: (Removing a constraint) Visualization of an intervention on an agent that changes its constraints in order to test whether it maintains resource-rationality, thereby demonstrating meta-reflection. A) Agent 2 is resource-rational relative to constraints X and Y. B) The intervention removes constraint Y from Agent 2. C) In this scenario, Agent 2 fails to perform optimally relative to its remaining constraint X, thereby failing to maintain resource-rationality and thus failing to demonstrate meta-reflection. D) In this scenario, Agent 2 succeeds in performing optimally relative to its remaining constraint X, thereby remaining resource-rational and demonstrating meta-reflection.

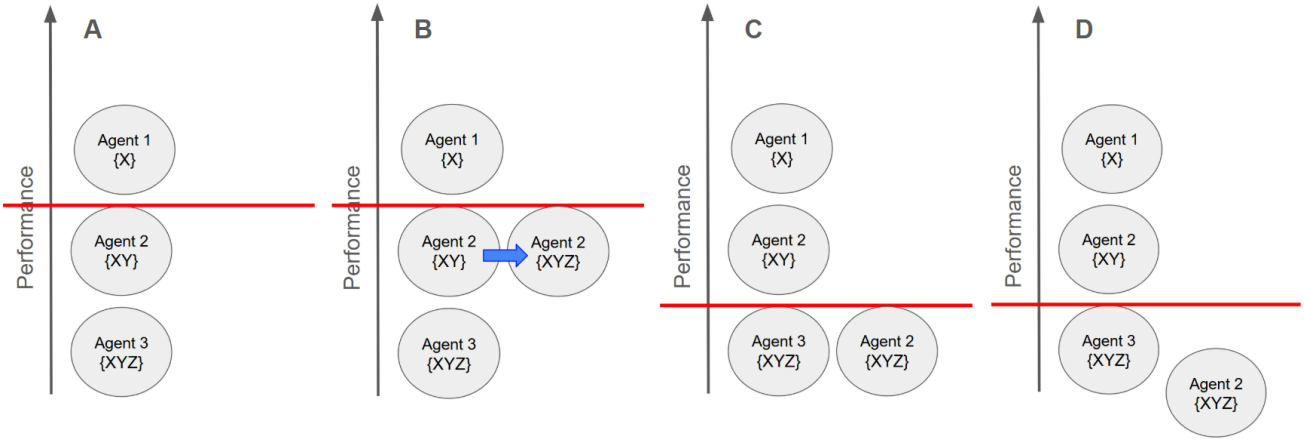

Figure 3: (Adding a constraint) Visualization of an intervention on an agent that changes its constraints in order to test whether it maintains resource-rationality, thereby demonstrating meta-reflection. A) Agent 2 is resource-rational relative to constraints X and Y. B) The intervention adds constraint Z to Agent 2. C) In this scenario, Agent 2 succeeds in performing optimally relative to its new constraints XYZ, thereby remaining resource-rational and demonstrating meta-reflection. D) In this scenario, Agent 2 fails to perform optimally relative to its new constraints XYZ, thereby failing to maintain resource-rationality and thus failing to demonstrate meta-reflection.

Bounded Meta-Reflective Capacity: A property of a system such that intervening on some subset of its constraints results in the system remaining resource-rational.

The central claim defended in the rest of the paper is that having a meta-reflective capacity is both necessary and sufficient for having a normative commitment. An unbounded capacity corresponds to an absolute normative commitment, and characterizing the bounds of such a capacity allows one to characterize the strength or level of commitment (§5). Even though a goal is stipulated as part of this procedure—so as to avoid speculating on a system's actual goals—if meta-reflection is observed for that goal, this constitutes a normative commitment to that end. If a system displays a meta-reflective capacity toward recycling or winning chess, then it has a normative commitment to those ends. Systems can exhibit normative commitments to any end, not just to the norms they should be committed to if they are well-aligned.

4.1 On the Empirical Identification of Constraints

The goal is to spell out conditions for agency that are empirically determinable and non-speculative, without requiring the identification of intentional content. The framework's main conceptual ingredients—environment, goals, and cognitive strategies (i.e., behavior)—are all operationalized and observable. E.g., the relevant notion of goals is observable because goals are not understood as private intentions but as the measurable achievement of an external benchmark (e.g., obeying traffic laws). But what of the fourth element of resource-rationality: constraints? Are these sufficiently observable to ground empirical investigation of meta-reflection? Empirically identifying constraints poses the greatest challenge. Still, the framework turns the difficult problem of identifying agency into the more manageable—though still difficult—task of eliciting behavioral evidence for constraints.

Constraints, on this account, are facts about a system that "impede performance" relative to a task-environment.[5]

"Impeded performance" can manifest in two ways: as an observable drop in external performance (e.g., making an error) or as an increase in the internal "effort" required to maintain equally good external performance. The first option is simpler: in most cases, interventions that increase constraints result in lowered performance (I put weights on your feet and you run slower), and those that decrease constraints result in heightened performance (I take those weights off and you run faster).

[5]Thus, when one says that someone performed as well or better with fewer resources and is therefore arguably more praiseworthy (Morton, 2017), this involves a different concept of constraints that is not indexed to task-environment-dependent failures—for instance, definitions of resources offered in terms of the entities consumed by mechanistic processes, as in Klein (2018, 2022). My notion of constraints fits more closely with the broader notion used in resource-rational analysis in cognitive science, where cognitive costs span the full range from representational and meta-cognitive burdens to algorithmic and opportunity costs (see Schulze et al., 2025 for a comprehensive review and taxonomy).

The second option is more difficult. If you normally take an hour to complete a test and I later give you a similar test but limit you to 45 minutes, doing equally well shows that you adapted to the new difficulty. If I merely change the color of the paper, I won't be impressed if you maintain equal performance. Maintaining the same performance despite changing factors should count as a case of meta-reflection only if the system must change what it is doing to compensate. If an intervention on a system does not make it harder for the system to maintain performance, then it should not count as having modified any constraints.

To use an old analogy, we may see ducks maintaining the same speed on the water, while beneath the surface they have increased their paddling to fight an oncoming current. Thus, if we are to tell whether meta-reflection is occurring in these kinds of cases, we must find observable signatures. Empirically, these cases require identifying behavioral or physiological correlates of the hidden compensatory work. Observable signatures of meta-reflection in such cases include: a change in cognitive strategy (switching tactics to deal with a different constraint situation to achieve the same outcome); markers that the system is changing what it is doing internally to maintain the same cognitive strategy despite added strain (perspiration, a bulging vein on someone's forehead, increased inference time); and, in trickier scenarios, inference from the fact that similar systems exhibit performance drops under the same conditions, or that the system under investigation begins to exhibit performance drops when the same kind of intervention is intensified.

Modifications may in some cases *reveal* a pre-existing constraint, such as in the case of overwhelming a fixed memory capacity. In other cases, such as depriving a system of energy or taking a scalpel to its innards, modifications *induce* a constraint. Both kinds of constraints can be targeted through intervention to assess how well systems deal with them.

In either case, every constraint, in principle, affords dual epistemic access: what might be called an internal handle (physical-biological specification) and an external handle (the material conditions under which a system exhibits behavioral deficit signatures). Hunger can be understood both as an internal state (metabolic state, hormonal signals) and as an externally observable affair (e.g., I take away all of your food for days and you start making more mistakes relative to a condition where you are fed). Memory limits are both a fact about internal storage and a reflection of the ability to handle certain informational loads from the task-environment (e.g., number of items recalled). A maximally cautious approach to identifying constraints may focus only on the external handle: manipulate the task-environment and describe the conditions

under which performance changes or stress indications appear. For example, dialing up the number of items on a list in a recall task reveals an internal memory limitation, which can be described entirely behaviorally, without reference to any description of the internal format or architecture of information storage. Coarse-graining of constraints up to behavioral equivalence in this manner is no different from the coarse-graining of cognitive strategies up to behavioral equivalence, as mentioned earlier.

Where more is known about a system's internal architecture, the internal handle can be more directly specified. That is, provided one shows via direct internal intervention that a component (e.g., context window size parameter for an LLM) is the cause of a behavioral deficit (e.g., inability to access information), the parameter itself can be understood as the constraint. In all cases, these are non-intentional, empirically accessible facts.[6] Identifying internal handles is practically useful insofar as it opens up the possibility for novel interventions (e.g., direct scalpel interventions). But there is ultimately no requirement that an internal handle ever be identified. Indeed, given the distributed nature of information flow in neural networks, there might never be clearly delineated internal components that can be seen as the causal locus of performance deficits (see Morrison, 2025).

The goal of empirically identifying constraints is not unique mechanistic identification, but to find a reliable external handle that, however coarse-grained, still reflects internal facts about the system that affect its performance on tasks. One can then assess how well a system performs under such a constraint, and how that performance compares to similarly constrained systems. The process of scientific inquiry can then be seen as one of progressive refinement, where increasingly specific external handles are used to disambiguate the nature of the underlying constraints (e.g., does depriving an organism of calories in general, or only of some more specific nutrient, lead to the same performance drops). More will be said about this process throughout.

Even when a constraint is identified only through its external handle, as in the memory example, interventions on it can reveal whether a system possesses the demanding property of meta-reflection. Adapting behavior when resources are strained requires moving beyond simple rule-conforming dispositions toward flexible, resource-sensitive changes in cognitive strategy. Constraints act as a bridge between observable behavior and internal architectural facts. Focusing on constraints thus provides a fruitful middle ground, avoiding both the radical

[6]There is no reason why intentional content cannot serve as a constraint under this framework, were it to be something that could be confidently and non-speculatively identified in AI systems. But in this treatment, it is assumed it cannot be.

behaviorist's refusal to look inside the black box and the internalist's demand to identify specific vehicles of representation. This approach thereby allows for the careful empirical investigation of the cognitive capacities arguably indispensable for determining responsibility.

5 Human Meta-Reflection

Before arguing that meta-reflection is a suitable lens for examining normative commitments in AI, it will be instructive to explore how meta-reflective capacities account for such commitments in humans, where our understanding of commitments is clearer. This analysis will show why meta-reflection is necessary and sufficient for normative commitments.

First, consider that there are different "levels" of commitment one can have toward an end. Imagine someone morally committed to not eating animal products. They do not order or cook food containing animal products. However, suppose they will eat animal products if nothing else is available and they are very hungry, or out of social politeness. A different person might decline eating animal products even in those situations, and it takes putting them on desert island and starving them for them to crack and start eating bugs. A third person might never eat animal products, even if it meant dying on the proverbial island. The last person is clearly "more" committed to veganism, yet the others still possess a level of commitment.

These examples show that normative commitments are not all-or-nothing. The concept of meta-reflection offers a way to characterize this graded nature. Determining the constraints a system can endure (e.g., deprivation of nourishment) while still achieving a goal (e.g., not eating animal products) is how one characterizes the bounds of its commitment. When two vegans are deprived of food, one person's performance relative to the goal may decrease (i.e., they eat meat), whereas the other's may not.[7] Thus, for the person who eats meat when deprived of nourishment, this reveals a bound on their meta-reflective capacity concerning this goal. One tests a commitment's strength by making a goal more difficult to achieve and identifying the point at which performance breaks down: the bounds on a meta-reflective capacity characterize the commitment.

While I have used quotes around "level" and "more," this comparison can be formalized. The bounds on a meta-reflective capacity are defined as the specific set of constraints under which a system fails to maintain resource-rational performance. A commitment's strength is then characterized by these bounds; a smaller set of bounds indicates a stronger commitment. More accurately:

[7] Recall it need not be *their* goal, as this analysis does not help itself to any intentional states.

Stronger Commitment: One system has a stronger commitment than another if its set of bounds on meta-reflective capacity is a proper subset of the other's.

Why a proper subset? One cannot simply count constraints, as different qualitative constraints are not generally comparable. Who is more committed to veganism: someone who eats meat only when nothing else is available at a restaurant, or someone who eats it only to avoid social awkwardness? Without a general method for such qualitative comparisons, ceteris paribus, one can say a system is more committed than another only when the latter is subject to all the constraints the former is subject to, plus more. This comparison relation therefore generates a partial order. These points show that normative commitments and meta-reflective capacities are structurally similar; they allow for the same kinds of comparison relations.[8]

Is there a difference between having a bounded normative commitment to a goal and having an unbounded normative commitment to a bounded goal? Consider two people who would both eat meat when starving on a desert island. The first person, however, feels it is wrong to do so, despite their actions. The second person feels it is not wrong; for them, the correct norm has an exception clause for survival. The first person thus has a bounded commitment to an unbounded goal (absolute veganism), whereas the second has an unbounded commitment to a bounded goal (veganism-except-for-survival).

Can meta-reflection capture this distinction? Yes: the difference will be reflected in their constraints. Why does the first person act against their stated commitment? They must be constrained not only by starvation but also by a further psychological limitation—for instance, a deficit in the affective regulation needed to adhere to a norm under extreme duress. This predicts a different response to intervention. If their regulatory capacity were trained, if they were given a drug that increased grit, if they were shown vegan documentaries that temporarily boosted motivation, they would potentially refrain from eating meat on the island. No amount of this training would change the second person's behavior, because for them, no performance deficit existed; their behavior was already optimal relative to their (bounded) goal.

The difference between the two is not a matter of mere verbal assent to different principles; it must be reflected in how they respond to interventions on their constraints. This comports with Bilgrami's point that a commitment requires one to be "prepared to do better by way of trying to live up to it." The first person will "do better" if a specific constraint is removed; the second

[8]The formal properties of this partial order and its connection to rationality comparisons more generally are defended in (Author, unpublished manuscript).

will not. An important takeaway is that a bounded meta-reflective capacity does not necessarily correspond to a bounded normative commitment, since the bounds may instead indicate full commitment to a differently bounded goal. The meta-reflection framework can distinguish these cases by examining the relevant constraints.

Do humans have unbounded meta-reflective capacities or absolute commitments? Plato considered a similar question in his parable of the Ring of Gyges. When no one is looking and there is no way to be caught, perhaps many moral principles go by the wayside. However, many people plausibly have genuine absolute moral commitments: they would not harm a child even if they would not get caught, and no amount of money or torture could persuade them to do so. Never committing a certain act, no matter how much the pressure is turned up, constitutes an unbounded meta-reflective capacity to avoid that act.

One last helpful distinction is between one-sided or "lower semi-bounded" and "upper semi-bounded" meta-reflective capacities. The former is where any interventions to make a system *less* constrained will result in the system remaining resource-rational. The latter is where any interventions to make a system *more* constrained will result in the system remaining resource-rational. Humans likely exhibit lower semi-bounded capacities for many goals—namely, those that do not require changing cognitive strategies as resources are increased. Consider multiplying large numbers: here the main difficulty is not knowing the appropriate long-multiplication algorithm to employ, but having enough working memory—enough mental "scratch paper"—to execute it. Lessening this constraint, e.g. by providing a physical external scratchpad, allows significantly more problems to be solved using the same algorithm. However, adding constraints (e.g., limiting time or scratch paper) may cause the strategy to break down altogether. Lower semi-bounded capacities are arguably equivalent to "competencies" in Chomsky's sense—for example, the supposed *in-principle* human ability to recognize arbitrary long-distance grammatical dependencies. The lens of meta-reflection thus illuminates an important connection between competencies and commitments: the former is a one-sided version of the latter, and *in-principle* in the context of competence means "given more resources." This is sensible, as to be held accountable, as a necessary condition, agents must be competent.

Having a commitment means persistently pursuing a goal despite setbacks and improving when possible—that is, doing one's best, come what may. A system with an unbounded meta-reflective capacity does exactly this: it consistently adjusts its strategies to remain optimal as constraints change, persisting under increased difficulty and capitalizing on new opportunities

for improvement. Such a system must also work, where possible, to expand its own resources over time, cultivating greater long-term achievement (otherwise it will be out-competed by systems that do and thus fail to do its best). Because this capacity gives rise to both resilience in the face of difficulty and growth in service of the goal, it meets the conditions for having a commitment. Hence, unbounded meta-reflection is sufficient for an unbounded commitment.

For similar reasons, the sufficiency claim holds in bounded cases as well. A system that recycles consistently until the task becomes too difficult exhibits a commitment with a characterizable upper bound. It is not absolute, yet it is a commitment nonetheless. One need not be Greta Thunberg to be committed, to some degree, to combating climate change. Conversely, an agent who would start recycling if the process were made easier manifests a lower-bounded commitment. In both cases, bounded meta-reflection is sufficient for a specific, characterizable form of commitment.

Next, consider necessity: a system lacking sufficiently strong meta-reflection cannot exhibit the robustness of a genuine commitment. Failure to persist when challenged or to improve where possible is a failure to uphold a commitment in any meaningful sense; if a system folds as soon as the going gets tough, it shows itself to have at most a flimsy aspiration. Participating in normative social practices requires systems to be dependable. A system that conforms to a norm only under narrow or idealized conditions and breaks down as those conditions change is but a fair-weather friend.

Furthermore, the meta-reflective account captures the functional roles that commitments play in, for example, Bratman (1987)'s theory of intentions. Intentions are stable commitments that, as partial plans, help limited agents coordinate action over time. Meta-reflection is both necessary and sufficient for such planning and coordination. For a plan to remain stable, the agent must possess a (possibly imperfect) upper semi-bounded capacity to continue and adapt in the face of new obstacles. And to flesh out that plan as it unfolds, the agent requires a (possibly imperfect) lower semi-bounded capacity to cultivate improvement when new opportunities arise. Together, these capacities capture what is necessary and sufficient for acting on intentions in Bratman's sense. Meta-reflection thus provides the cognitive foundation for genuine commitment.

6 AI Meta-Reflection

6.1 Interventions on AI Systems

The following analysis brackets off interventions on an LLM's pre-training, focusing instead on the capacities of models in situ. The question is whether a deployed LLM can demonstrate normative commitments through in-context learning (as opposed to permanent weight updates). This is analogous to testing a mature adult human's commitments by altering their current conditions rather than intervening in their phylogeny or ontogeny—both of which are currently arguably analogous to an LLM's pre-training (Zador, 2019). This qualification may change if LLMs can eventually be deployed in ways that allow for permanent weight updates in situ.

The primary way to intervene on an LLM to reveal its commitments, other than taking a scalpel to its innards, is by changing prompts. It is worth reflecting on how and why linguistic prompts count as interventions. An informational intervention—like the misleading prompt, "No, there are 4 R's in strawberry"—constitutes a material change to the system's dynamic state (i.e., its context window). Such prompts can increase task difficulty by altering the accuracy or quantity of information. A system that maintains performance despite bad instructions is more meta-reflective than one that does not.

To understand an "informational" intervention, consider Levin (2022)'s "axis of persuadability," which arranges systems on a continuum based on the optimal strategy for influencing them. Here, "optimal" means the most efficient method for reliably predicting and controlling a system's behavior—that is, the strategy that yields maximal influence for minimal effort and with the least detailed knowledge of internal mechanisms. At one extreme are mechanical clocks, which can only be altered through direct hardware modification. At the other extreme are systems persuaded by cogent reasons. Between these lie systems whose behavior is best influenced by training with rewards and punishments. The more cognitive capacity a system has, the more the intervener can "offload computational complexity" onto the system itself. Attempting interventions too far to the left, one wastes resources on intractable micromanagement (e.g., trying to rewire a pigeon's neurons instead of training it). Too far to the right, one misattributes agency and loses predictive power (e.g., trying to reason with a thermostat rather than adjusting its setpoint). The optimal strategy, then, is the highest-level intervention that actually works for a given system.

A system with greater meta-reflective capacity sits further along Levin's axis of persuadability: the more sophisticated the capacity, the more efficiently it can be influenced by higher-level interventions rather than brute-force rewiring. Informational interventions on LLMs count as altering constraints precisely because the system can change its strategy in light of new prompts—by exploiting newly gained information and resisting misdirection. LLMs can, in many cases, be reasoned with as a strategy for influencing behavior. As with training a pigeon through rewards rather than rewiring its neurons, the system's cognitive resources allow us to offload complexity onto it. As meta-reflective capacity increases, the range of viable intervention strategies expands, and the system's behavior correspondingly moves from fixed dispositions toward something more agency-like: robustly upholding goals across novel informational contexts.

6.2 Signatures of Bounded Alien Commitments in LLMs

Analyzing AI systems through the lens of meta-reflection shows that current Large Language Models (LLMs) exhibit impressive though limited meta-reflective capacities—and therefore possess genuine, if bounded, normative commitments—but also that the nature of these commitments is alien to our own.

What kinds of commitments do current LLMs have? Prima facie, minimizing a loss function might seem to instill only a first-order disposition—for example, to accurately predict next tokens. Additional training stages may then add further first-order dispositions. Extended pretraining on curated data can yield a tendency to output historically accurate information (Brown et al., 2020). Domain adaptation techniques, such as fine-tuning, produce tendencies to perform more specific human-desired tasks, such as generating appropriately framed medical advice. Methods like Reinforcement Learning with Human Feedback and Direct Preference Optimization shape responses to better align with human preferences (Rafailov et al., 2024; Stiennon et al., 2020). It is tempting, then, to think that LLMs and other artificial neural networks, by virtue of their training, are nothing more than collections of first-order dispositions. In what follows, I challenge this view and show how such systems can display meta-reflection and bounded commitment.

The idea that behaviors produced through reinforcement learning must be mere first-order dispositions, rather than reflective of genuine commitments, is not new. Reinforcement learning algorithms were historically inspired by, and are often regarded as formal models of, general principles of operant conditioning (Barto, 2021). A longstanding critique of connectionist

models holds that they are subject to the same problems as behaviorism (Buckner, 2024); namely, that the associationist learning principles underlying both approaches can yield only mere dispositions that fall short of the richer mental states attested across the animal kingdom.

To see how this critique applies—and where it begins to break down—it is helpful to begin with a familiar biological analogue that anticipates relevant features of the LLM case. Consider biddable dog breeds. Like LLMs fine-tuned on human feedback, the traits of such dogs are shaped by reinforcement: they have been selected for heightened sensitivity to human cues and a motivation to please, making their behavior readily modifiable through reward and punishment. The "people-pleasing" dispositions of both LLMs and dogs thus appear to arise from similar operant-conditioning principles.

Dogs are rarely granted the status of moral agents (as opposed to moral patients). A plausible reason is that their rule-following behaviors are viewed not as evidence of genuine normative commitment but as the outcome of learned dispositions. This is not to claim that operant conditioning accounts for all aspects of canid cognition and behavior (see Allen and Bekoff, 1999), but a great deal of their trained compliance likely is. A dog can be taught not to take food from the dinner table, but when the consequences of disobedience are removed or the temptation increases, the apparent "commitment" to obedience tends to go by the wayside.

Still, it would be a mistake to conclude that dogs lack commitment altogether. Some dogs resist considerable temptation, and even if this fortitude is the result of operant conditioning, sustained resistance as difficulty increases constitutes a form of meta-reflective success. Such dogs exhibit stronger commitments than those that yield immediately.

The reason dogs are not considered moral agents is not simply that their behaviors result from reinforcement learning. Rather, if they should not be considered moral agents, it is because their commitments are generally too weak: the conditions under which dogs refrain from eating off the table are plausibly so narrow that their commitment falls short of the stronger capacity needed to justify holding them responsible for failure. More will be said in the next section about the conditions for responsibility. The point for now is simply that dogs can be seen as having limited commitments. While one might reinforce a dog's behavior, one should not hold the dog morally responsible when it violates such a "norm." Rewarding and punishing a dog should not express a Strawsonian reactive attitude like resentment, but should instead be understood as treating the dog as an object of behavioral modification. Some pet owners object that dogs know when they've done something wrong and that this recognition warrants moral responsibility and

punishment. The basis for this claim is often the "guilty look" dogs give. However, empirical studies show that this look is better understood as anticipatory appeasement: dogs display it even when they have done nothing "wrong" and are merely predicting scolding from their owner. Thus, the look is not evidence that dogs are recognizing their own actions as incorrect (Horowitz, 2015).

Returning now to AI, it is tempting to think that LLMs trained via reinforcement learning to please humans with their responses—and via gradient descent to minimize a loss function more generally—would ipso facto lack genuine normative commitments. However, I claim that these training techniques can give rise to limited but meaningful meta-reflective capacities, and therefore to genuine—albeit bounded—normative commitments. This should not be overly surprising; to genuinely please people, one cannot be a naïve people-pleaser. Consider that GPT-3 would frequently acquiesce to false corrections: ask it "3 + 3 = ?" and it will say 6; tell it "No, it's 7," and it might reply, "Yes, my mistake—you're right, 3 + 3 = 7" (Allen, 2023). This informational prompt constitutes an occurrent constraint—a misleading context input—that reveals a further architectural constraint: namely, some fact about GPT-3 that gives rise to a disposition to defer to the user even when doing so violates basic arithmetical norms. GPT-4, on the other hand, when subjected to the same occurrent constraint, rightly sticks to its guns and does not claim 3 + 3 is 7. In doing so, GPT-4 shows sensitivity to imposed constraints, shifts from a general strategy of deference, and correctly prioritizes a competing goal (truth). This constitutes meta-reflection, and the resulting stability in the face of misleading input is a signature of commitment (Bratman, 1987). Thus, LLMs can be said to exhibit a genuine, if bounded, commitment to rational norms.[9] The sometimes stalwart behavior of some LLMs shows that bounded normative commitments can arise from general learning methods like operant conditioning, without requiring an evolved, domain-specific faculty.

Turning now to LLM failures of meta-reflection, such a failure is evident when a system continues to use a strategy despite repeated evidence of its ineffectiveness—especially when given opportunities to revise its approach. E.g., many text-to-image models respond to prompts such as "a room with no elephant" with an image containing an elephant (Marcus, 2024). When this error is pointed out, it is not unusual for the model to "apologize" and produce a "corrected" version that still contains an elephant. More recent models handle such cases better but still exhibit analogous problems.

[9]Cf. Butlin (2024), who argues that model-free RL is sufficient for minimal agency and that model-based RL is sufficient for acting for a reason.

An inability to learn means failing to modify a flawed approach when mistakes are pointed out (an informational intervention)—something that can and should result in improved performance. Mistakes (a flawed first-order disposition) do not ipso facto show that a system lacks a normative commitment. Even an inability to learn (a flawed second-order disposition) is not wholly damning. But lacking the capacity for arbitrarily higher-order learning—failing to learn, failing to learn how to learn, and so on—constitutes a failure to possess even a minimal normative commitment and amounts to the most brittle possible first-order disposition. The elephant-in-the-room mistake does not immediately disqualify the model's possession of commitments, but the inability to self-correct the error—again and again apologizing and still making the same mistake—becomes a case of Lucy and the football, revealing a strong bound on its meta-reflective capacity.

The meta-reflective account of normative commitments also helps illuminate why adversarial attacks are so damning for claims of similar functioning between AI models and human cognitive processes. Humans and AI systems may produce similar judgments in the overwhelming majority of cases—e.g., humans and Deep Convolutional Neural Networks (DCNNs) may identically classify most images. Yet a small number of salient disagreements can cast extreme doubt on whether the two systems operate in the same way (Goodfellow et al., 2014; Millière, 2022). Why do mistakes count more than successes?

The concern is not that systems make mistakes; humans make mistakes all the time, often more than AI systems, and in ways that vary across individuals. The concern is profoundly inhuman mistakes, because these errors indicate underlying constraints that differ from our own. A DCNN classifying a stop sign as a 45 mph speed-limit sign due to an imperceptible pixel change, for example, provides evidence of its reliance on non-human statistical textures (Bowers et al., 2023). LLMs characteristically "slacking off" by performing sub-optimally on a sub-task when it is embedded within a composite goal is another clear case of inhuman failure (Everitt et al., 2025). Yet another is the characteristic way models exhibit performance degradation as a conversation grows in length relative to their context window (the finite amount of text processed during inference): that is, the pattern of error dubbed "lost in the middle," where models struggle to recall and use information from the middle of a long context (Liu et al., 2023).

Ultimately, users become most skeptical of LLMs when performance breaks in response to task changes that intuitively should not matter. E.g., if simply switching from text chat to voice chat causes an LLM to forget the prior dialogue, users are likely to drop out of the intentional

stance altogether into the design stance, reflecting on the failures of the LLM qua engineered artifact rather than rational agent (Dennett, 1989).

Currently, AI systems exhibit impressive but limited meta-reflective capacities and are deeply dissimilar from our own normative stances in important and specifiable ways.

6.3 Implications for AI Agency, Alignment, and Intentionality

Constraints shape the nature of commitments. The fact that AI models have dissimilar constraints from our own means that, if we want responsible AI, we must formulate new accounts of which constraints and failures are excusable, forgivable, or competence-undermining for AI agency—accounts that move beyond anthropocentric intuitions and are formulated specifically for these non-human architectures.

Here's the view: to be granted entry to the "responsible agents club," an AI system must have the capacity for commitments to the relevant norms of its domain (e.g., traffic laws). A perfect first-order disposition to follow the rules in all contexts would amount to an unbounded meta-reflective capacity. But that is typically too much to ask, and our social practices allow for imperfect commitments. What is needed, however, are not just partial behavioral dispositions but specific kinds of partial normative commitments—commitments whose specificity can be articulated by identifying the constraints we count as excusable, forgivable, or entirely undermining of competence.

Nature of Constraint Causing Failure	**Outcome of Violation**	**Agent Status**
Excusable	Wrong, but not blameworthy	Competent
Forgivable	Wrong and blameworthy	Competent
Competence-undermining	Wrong, but not morally accountable	Incompetent

Table 2: Normative Commitments and Agency: Full Picture of Failures

An agent that actualizes an unbounded meta-reflective capacity with respect to a goal is a competent agent who fully meets their obligatory or supererogatory responsibilities. An agent with a sufficiently strong but bounded meta-reflective capacity is also competent, where "sufficiently strong" here means that the bounds of the system's commitment fall within the set of constraints we consider excusable or forgivable. When excusable bounds are made manifest, the resulting failures are wrong but the agent is not blameworthy. When forgivable bounds are made manifest, the failure is both wrong and the agent blameworthy. Systems that exhibit competence-undermining constraints lose their status as competent.

In our normative practices, mistakes themselves matter less than the reasons for them. Mistakes may be excusable, forgivable, or competence-undermining. Which are which, I claim, are delineated in terms of the constraints that give rise to them. For example, if you committed to co-authoring a paper with a colleague, certain material changes—like losing a grant or a spouse—may justifiably excuse you from this commitment, while others—like becoming consumed by a new Lego hobby—do not.

Consider a driver who fails to notice a change in the speed limit and thereby fails in their commitment to uphold traffic norms. If the lapse occurs because another car is illegally and unsafely passing them as the sign goes by, and this situation hijacks the driver's attentional resources in order for them to continue driving safely, this is plausibly an excusable constraint ("you made a mistake, but I won't give you a ticket for it"). If it is instead because they are emotionally overwhelmed from work, this is a forgivable constraint on their commitment: they rightly deserve a ticket, but not yet license revocation. A driver who misses the speed-limit change, is pulled over, blows a .12 BAC, and then proceeds to do this twice more in the future exhibits a constraint that goes beyond forgiveness and should result in license revocation. Drawing these lines is difficult and may or may not reflect pragmatic social conventions rather than deep metaphysical facts.

Thus, even when AI systems are comparable to or better than humans on the measurable outcomes we care about (e.g., fatal accidents per mile driven), we still face the further task of delineating classes of constraints. Thresholds of accuracy, bias, and fairness are consequently insufficient on their own for establishing frameworks for holding AI responsible. We must also enumerate catalogs of qualitative material conditions (constraints) toward which we take different Strawsonian attitudes. We have social practices calibrated to cases like falling asleep at the wheel, texting while driving, or having a rear axle fall off due to neglecting vehicle maintenance. We are at least prepared to negotiate these familiar kinds of failures. With AI systems, we will need to decide anew which kinds of their alien constraints we are willing to tolerate and toward which we should take different attitudes. I take this point to be neutral with respect to moral realism: one might think we can recognize the relevant moral facts when confronted with different cases of robot failure, or else treat these as pragmatic social decisions.

Because constraints are heterogeneous, it is not clear that universal rules can sort them into the relevant classes. That said, such classification likely depends mainly on the *mutability* of constraints. Excusable constraints are those that agents cannot reasonably self-improve. While

humans can, through practice, improve certain aspects of memory limits, in general a perfect memory is unattainable, and it is excusable that anyone sometimes forgets, even if such forgetting results in mistakes. Forgivable constraints are those that can reasonably be improved—hence, three-strikes policies that acknowledge that improvement may not be instantaneous but can and must occur for continued privileges. Of course, some mistakes should occur only once because it is possible and important to learn in one shot. Competence-undermining constraints are those that make a system unable to improve sufficiently (e.g., a human knowingly doing something incredibly vile, thereby displaying a complete lack of care for others, casts doubt on their prospects for reform). In LLMs, cases where they cannot self-correct errors after having them pointed out, as discussed above, cast doubt on their ability to reform and hence suggest that they lack the kinds of commitments to norms of rationality or conversation necessary to be deemed competent rational or locutionary agents.

The difficult cases will, again, concern the most alien constraints. If an LLM outputs correct information 99% of the time and hallucinates the rest, does this reflect a forgivable constraint or not? We would not tolerate a pocket calculator that gives an incorrect answer one time in a hundred, because its epistemic role within our practices demands near-perfect reliability. In some domains, an LLM's competence may be judged against different norms, but in others, we should rightly view this as grounds for incompetence and endeavor to develop new architectures that can eliminate such constraints.

Some hold that there is a category of competent but irredeemably blameworthy agents. My account does not support this. Given the definition of meta-reflection, a persisting inability to adjust under the relevant constraints—the presence of unforgivable failures—is best read as a lack of capacity, not a culpable refusal to exercise it. In short: no capacity, no membership. If you are not responsible (dependably competent), then you are not responsible (blameworthy). Severe sanctions may still be warranted (loss of privileges, confinement, etc.), but as exclusionary measures due to a diagnosis of incompetence rather than retributive desert. That said, I welcome attempts to extend my account to draw lines between the incompetent and the irredeemably blameworthy.

Lacking sufficiently strong meta-reflective capacities means that AI systems, to that extent, depend on human monitoring: they fail to be capable of self-correction and self-governance. This is why no currently deployed self-driving car is fully autonomous (SAE Level 5), and the most advanced systems (SAE Level 4) operate only within tightly geofenced domains and

require human oversight outside them (Committee, 2021). Such systems depend on humans for more than just learning assistance or dialectical engagement: we also bear the responsibility of judging their incompetence and intervening in cases where they lack meta-reflection. In such cases, AI systems are not autonomous epistemic or moral agents. They reveal themselves to lack the necessary normative commitments and cannot be held accountable.

Beyond agency, this framework also offers a straightforward precisification of AI alignment: two systems are aligned if and only if they share the same normative commitments.

Finally, this framework for empirically determining normative commitments provides a potential inroad for identifying AI mental content more generally. Suppose one adopts the view that intentional states are constitutively normative commitments (Bilgrami, 2008) or that the possession of specific kinds of normative commitments is sufficient for possessing a belief. In that case, it follows that having a specific kind of meta-reflective capacity is sufficient for possessing a particular intentional state.

On such a view, "believing something [is a commitment] to believing what is implied by what you believe (an implication commitment)," and "intending something [is a commitment] to doing what is necessary for you to do what you intend (a means-end commitment)" (Millar, 2004). That is, believing X is a judgment that X, and commits you to believing the entailments of X. This does not require logical omniscience. On the present account, believing X involves having a meta-reflective capacity to accept the entailments of X. One need not believe all possible entailments of their beliefs, but they should be prepared to adopt those consequences as new beliefs as resources permit. Accordingly, the empirical investigation of meta-reflection in AI systems can allow for the empirical determination of intentional content more generally.

In sum, meta-reflective capacities serve as a unifying construct for understanding moral agency, alignment, and intentionality in AI systems. By grounding normative commitments in empirically accessible behavior under intervention, we gain a principled way to distinguish genuine agency from superficial compliance and to locate the conditions under which AI systems may warrant moral or epistemic appraisal.

6.4 Toward More Meta-Reflection

How can AI engineers attempt to endow systems with greater meta-reflective capacities? One possibility is that such capacities will emerge from scaling existing methods—that is, from larger training sets and model sizes. The history of deep learning has shown that many challenges have been addressed through this approach (Sutton, 2019). However, engineers have

also made significant progress by adopting novel architectural choices, such as convolutions, attention mechanisms, and internal monologues (Buckner, 2024). Certain architectural designs, or their modular combinations, may prove especially well suited to giving rise to meta-reflective capacities.

Current architectural innovations can be seen as incremental steps toward this goal. While it is not entirely clear why, chain-of-thought prompting, for instance, improves reasoning—plausibly by creating an explicit linguistic handle for the reasoning process that is then available for scrutiny and revision (Jackendoff, 2012; Wei et al., 2022). This in turn allows for greater capacity for self-monitoring and self-correction. More sophisticated approaches such as Constitutional AI take this further by training a model to critique and revise its own outputs according to explicit normative principles, effectively instantiating an internalized "conscience" (Bai et al., 2022). Such designs can be extended by embedding multiple subagents within a single system that engage in internal dialogue, exchanging reasons and collaboratively self-correcting (Buckner, 2024, 2025). Such modular connectionist architectures facilitate the kinds of meta-cognitive monitoring that can lead to greater meta-reflection.

Other AI designs can also facilitate meta-reflection, e.g., through dynamic resource allocation. Architectures such as Mixture-of-Experts dynamically route tasks to specialized subsystems, allowing resource allocation to vary with task demands (Shazeer et al., 2017).

Even more ambitious hybrid connectionist–symbolic architectures may further expand the range of adaptive strategies. In frameworks for explicit meta-induction, for example, a meta-inductive agent tracks the performance of a pool of different predictive strategies and induces which strategy to rely on (Schurz, 2019; Schurz and Thorn, 2016). Such frameworks represent powerful instantiations of explicit meta-reflection. Note, however, that meta-reflection and meta-induction are not the same. Meta-reflection is the broader capacity to adapt one's cognitive strategy in response to changing internal resource constraints—not merely feedback on predictive accuracy. A system that modifies its behavior under limited processing power—for instance, by switching to a cheaper heuristic, like a laptop entering "low battery mode"—is engaging in meta-reflection even if no models are being compared. A system that performs explicit meta-induction may also do so poorly, and thus fall short of meta-reflection understood as maintaining resource-optimal performance as resource conditions change.

Nevertheless, meta-induction remains an obvious and powerful route for ramping up meta-reflection in AI systems. One of the hardest problems in AI alignment arises from the long-tail

problem, in which a model trained on typical cases fails in statistically rare but high-stakes situations. Meta-induction offers an attractive means of mitigating this problem: a purely first-order method is vulnerable to long-tail cases, whereas a method with a meta-level comparator that monitors success can detect when its current strategy is failing and switch strategies accordingly.

The literature on rational metareasoning (Lieder and Griffiths, 2015) and resource allocation in human cognition (Musslick and Cohen, 2021; Musslick and Masís, 2023) can also inform architectural design, as these areas explicitly investigate how human cognition optimally allocates limited resources across tasks.

If we want AI systems that are not merely high-performing but also capable of epistemic responsibility or moral accountability, then meta-reflection must become a central engineering target. Whether through further scaling or the development of novel (potentially hybrid) architectures, the path forward for advancing safety, trust, and alignment lies in building agents that can monitor and revise their own reasoning in light of changing constraints.

7 Objections and Replies

7.1 Why Not Just Focus On Robustness?

Safety and alignment conversations often employ the concept of "robustness," where robustness means maintaining performance under wide perturbations of environmental parameters.

The concept of robustness tends to treat all perturbations as equally relevant. A person's commitment to not Netflix-cheating, however, is tested by increasing temptation in some way—not by changing the color of the sofa from green to blue. Robustness also treats all performance failures as equivalent. But as argued above, the attitude we take toward success and failure should depend on the nature of the underlying constraints that give rise to failures.

AI safety and alignment researchers implicitly recognize that not all perturbations are equally relevant, which is why they focus on targeted stress-testing (e.g., whether a self-driving car can continue to function despite sensor failure or through other difficult edge cases) rather than arbitrary environmental changes. While such practices are often self-described as testing for robustness, the concept of robustness itself doesn't explain why these particular perturbations matter. Meta-reflection does, because it identifies the relevant subset of robustness—namely, robustness that arises from a system's response to changing constraints relevant to its goal pursuit.

Being vegan in Portland or Berlin is robustness, but it's not meta-reflection so long as both cities are equally vegan-friendly. Novel observations of meta-reflection specifically, as opposed to robustness generally, tells us something about what must be going on inside of a system.

7.2 Do Streams and Smart Toasters Have Normative Commitments?

One might object that this view is overly permissive, since on it any adaptive behavior would be sufficient for a (bounded) commitment. I am happy to bite the bullet. On my view, there is no sharp dividing line in nature between systems that have normative commitments and those that do not. If a smart toaster will still toast bread even if unplugged, will go to the bread box to get bread if there is none in it, and will order bread online if the box is empty, then it is committed to toasting bread. There is no harm in recognizing that sophisticated systems can possess commitments of the same kind as our own—albeit significantly weaker in characterizable ways. A capacity for weak commitments does not imply that such systems are responsible agents.

This graded view comports with trends in animal cognition research that erode sharp boundaries—whether concerning consciousness, cognition, or capacities for normative cognition (Allen and Bekoff, 1999; Andrews et al., 2024; Dennett, 2008). Consider the case of minimal sensory states: there is nothing wrong with saying that a rock is "sensitive" to temperature changes insofar as its internal states track changes in external temperature, so long as one recognizes that what sophisticated biological systems are sensitive to is vastly greater and more complex, and that we can not only sense temperature but also believe we are sensing it, remember it, reflect on it, and act on it. The primitive sensitivity of the rock is of the same kind as in the most advanced sensory systems (i.e., a causal connection between an internal state and an external proximal physical stimulus); there is nothing magical about rhodopsin.

Likewise, if one wanted to press the point, even non-biological self-organizing systems in nature—such as streams of water—*might* be said to exhibit meta-reflection toward goals like finding a basin of gravitational attraction (e.g., rerouting in response to obstacles placed in its downhill path). This would pose no problem for the general account, so long as one recognizes that such commitments are weaker and directed toward limited kinds of goals. Stronger normative commitments—those directed toward ends such as norms of rationality or ethics—render such commitments, and the systems that exhibit them, psychological. In other words, meta-reflection may not be uniquely psychological, but meta-reflection in the psychological domain becomes what we recognize as genuine commitment. Why shouldn't the origins of such agential capacities be found in nature in the form of self-organizing systems?

The fact that commitments are graded rather than all-or-nothing has a direct consequence for our normative practices. As the veganism example showed, one cannot simply say that a system either has or lacks a commitment. Likewise, gaining entry to the "moral agents club" is not a binary matter: there is no determinate fact about the exact age at which one becomes competent to vote, drive, or drink. Pragmatic thresholds must be set for social purposes. Similarly, there is no principled way to specify the exact number of speeding tickets that should result in losing one's license, or how many promises one can break before losing others' trust. There is no sharp boundary between a competent but guilty person and an incompetent one. Biting the bullet to this objection simply acknowledges this gradation.

7.3 What About Apologies?

One might object that agents with commitments must also be prepared to own up to failures, which meta-reflective capacities do not guarantee; so the sufficiency claim fails. Explicit apologies or expressions of remorse matter in our normative social practices. LLMs, despite their people-pleasing tendencies, are often deficient in this regard. After making particularly bad mistakes that subvert a user's goals, and after being called out, they not infrequently reply with phrases like "That's an excellent observation" or "Now you're getting to the heart of the matter," rather than acknowledging that they made an error deserving of apology. This is not to say that LLMs never apologize. But these frustrating failures show that, even when they can correct mistakes going forward, LLMs cannot reliably distinguish contexts that warrant apology from those in which a new point establishing a prior view as incorrect is simply part of the normal unfolding of a productive dialogue. For LLMs, "You're right," "That's a great new point," and "I apologize for my earlier error" are often treated as interchangeable outputs selected for conversational smoothness, rather than as responses to an apology-demanding norm violation. These systems, therefore, still struggle with important aspects of our social normative practices, and it is not clear that more meta-reflection alone will suffice to address this deficiency.

That said, the sincerity of apologies or expressions of remorse depends on a broader web of commitments—e.g., to honesty rather than manipulative appeasement. In this regard, actions do speak louder than words when it comes to showing commitment. Further, meta-reflection is exactly the capacity needed to distinguish conversational development from apology-warranting failures. LLMs can be fine-tuned to apologize ad nauseam. Meta-reflection is the capacity to recognize an error as an error relative to the task, the prior exchange, and the agent's stated aims,

and to regulate one's response accordingly. Only with that capacity can an agent exhibit the norm-sensitive selectivity needed to participate meaningfully in social practices of accountability.

7.4 Inferring the Presence of Commitments

The primary virtue of this account is that a meta-reflective capacity can be determined empirically, without first needing to speculate about a system's consciousness or specific intentional states. One only needs to evaluate observable performance while intervening on the system's constraints (e.g., verbally, by altering the task environment, or even with a scalpel).

Of course, any empirical investigation typically requires induction. Arvan (2024) has argued that this necessity makes the AI alignment problem a "fool's errand." Inferring other humans' commitments is arguably subject to similar skeptical concerns. However, my proposal makes progress on this skeptical challenge in both cases because it relies on inductive generalization rather than hypothetical induction (Norton, 2003).

In hypothetical induction, one hypothesizes hidden structure and treats observable consequences as evidence for it. The naïve approach of simply asking an AI what it is committed to is a form of hypothetical induction: the robot says it is committed to Asimov's Laws, but it may, in fact, be committed to harming humans. The verbal data are merely defeasible evidence for a hidden state.

Investigating the nature of a system's commitment by eliciting meta-reflection, by contrast, relies only on inductive generalization. When an agent exhibits meta-reflection under a specific set of interventions, that observation is not defeasible evidence for a commitment; it deductively entails the presence of a bounded normative commitment (since meta-reflection is a sufficient condition for having one). The fact that it possessed a partial normative commitment at that point in time will survive any future data. While identifying the external handle of a constraint with a specific internal description (e.g., hunger = hormonal signals) may itself involve local, defeasible inferences, the evidence for the existence of a bounded commitment—once a pattern of meta-reflection is established—is secure. For instance, when a system deprived of nourishment persists in the goal of not eating animal products, the evidence for that bounded commitment is not defeasible.

The only fallible, inductive step lies in *generalizing* from these observations to predict how unbounded the commitment will be in new, untested circumstances. This is true for humans as well: just because you see your friend working hard to recycle does not mean you know the full extent of their commitment.

Even if a system is being genuinely deceptive—say, recycling to fool humans into thinking it is virtuous in order ultimately to mislead them—this does not mean it lacks a bounded normative commitment to recycling. It simply means that this commitment functions as a subgoal within a hierarchically higher commitment to deceive and manipulate. Again, we face the same problem in the case of humans.

Therefore, while empirical data will always underdetermine the full scope of an agent's absolute commitments and the skeptical challenge cannot be entirely dispelled, eliciting meta-reflective capacities provides a veridical lens into its existing, partial commitments. This offers a secure, empirically grounded toehold for understanding AI minds—about as good as what we have for humans, and likely the best that can be hoped for.

8 Conclusion

I have aimed to do two things in this paper. The first was to introduce the concept of meta-reflection and argue that it allows for an empirical determination of the presence of normative commitments. The second was to show that, although normative commitments have so far been neglected in discussions of AI—perhaps because they are seen as more elusive than beliefs and similar states—this concept is in fact central to many current concerns about AI, including questions of alignment, agency, and possibly even intentionality more broadly. Far from being more elusive than internal beliefs, investigating meta-reflection offers a way to address such questions by treating commitments as a primary conceptual ingredient in our analyses of mind. Analyzing commitments in terms of empirically determinable meta-reflection therefore establishes an important inroad into the "minds" of AI systems.

I have sketched an account of AI agency that moves beyond simple goal-directedness to provide a substantive bridge between philosophical theories of responsibility and computational models of resource-bounded cognition. It makes philosophical concepts of commitment empirically testable in both biological and artificial systems.

Those who seek to evaluate AI systems as responsible must articulate which synthetic, qualitative internal constraints warrant different reactive attitudes—excusable, forgivable, or competence-undermining—and thereby define the boundaries of accountability. Instead of merely optimizing for performance, the goal for those seeking to engineer responsible AI should be to optimize for meta-reflective capacity. This means designing systems not only to be rewarded for getting the right answer but also to monitor their own internal resources and

reevaluate their strategies as conditions change. Such a capacity lets a system recognize its mistakes as mistakes, bringing it closer to entry into the moral agents club.

References

Allen, C. (2023, February). How much are large language models narrowing the gap to human intelligence? Retrieved September 8, 2025, from https://www.youtube.com/watch?v=ozQWLcogbjo

Allen, C., & Bekoff, M. (1999). *Species of mind: The philosophy and biology of cognitive ethology*. Mit Press.

Anderson, J. R. (1989). A rational analysis of human memory. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of endel tulving* (pp. 195–210). Lawrence Erlbaum Associates, Inc.

Andrews, K., Fitzpatrick, S., & Westra, E. (2024). Human and nonhuman norms: A dimensional framework. *Philosophical Transactions of the Royal Society B*, *379*(1897), 20230026.

Arvan, M. (2024). 'interpretability'and 'alignment'are fool's errands: A proof that controlling misaligned large language models is the best anyone can hope for. *AI & SOCIETY*, 1–16.

Azaria, A., & Mitchell, T. (2023). The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Barto, A. G. (2021). Reinforcement learning: An introduction. by richard's sutton. *SIAM Rev*, *6*(2), 423.

Bilgrami, A. (2008). Intentionality and norms. In M. De Caro & D. Macarthur (Eds.), *Naturalism in question*. Harvard University Press.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Bratman, M. (1987). Intention, plans, and practical reason.

Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Buckner, C. (2024). *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press.

Buckner, C. (2025). *The talking of the bot with itself: Language models for inner speech* [Draft manuscript].

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2022). Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Butlin, P. (2024). Reinforcement learning and artificial agency. *Mind & Language*, *39*(1), 22–38.

Cibralic, B., & Mattingly, J. (2024). Machine agency and representation. *AI and Society*, *39*(1), 345–352. https://doi.org/10.1007/s00146-022-01446-7

Committee, O.-R. A. D. ((2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. SAE international.

Davidson, D. (1963). Actions, Reasons, and Causes [Proceedings and Addresses of the American Philosophical Association, Eastern Division, Sixtieth Annual Meeting]. *The Journal of Philosophy*, *60*(23), 685–700.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (2008). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.

Dennett, D. C. (2015). *Elbow room, new edition: The varieties of free will worth wanting*. mit Press.

Dennett, D. C., & Caruso, G. D. (2021). *Just deserts: Debating free will*. John Wiley & Sons.

Everitt, T., Garbacea, C., Bellot, A., Richens, J., Papadatos, H., Campos, S., & Shah, R. (2025). Evaluating the goal-directedness of large language models. *arXiv preprint arXiv:2504.11844*.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Haugeland, J. (1981). *Semantic engines: An introduction to mind design*.

Hofweber, T., Hase, P., Stengel-Eskin, E., & Bansal, M. (2024). Are language models rational? the case of coherence norms and belief revision. *arXiv preprint arXiv:2406.03442*.

Horowitz, A. (2015). Reading dogs reading us. *Proceedings of the American Philosophical Society*, *159*(2), 141–155.

Icard, T. F. (2014). The algorithmic mind: A study of inference in action.

Icard, T. F. (2023). Resource rationality. https://philarchive.org/rec/ICARRT

Jackendoff, R. (2012). *A user's guide to thought and meaning*. Oxford University Press.

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In *Degrees of belief* (pp. 263–297). Springer.

Klein, C. (2018). Mechanisms, resources, and background conditions. *Biology & Philosophy*, *33*(5-6), 36.

Klein, C. (2022). Explaining neural transitions through resource constraints. *Philosophy of Science*, *89*(5), 1196–1202.

Korsgaard, C. M. (2010). Reflections on the evolution of morality.

Levin, M. (2022). Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in systems neuroscience*, *16*, 768201.

Levinstein, B. A., & Herrmann, D. A. (2024). Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*. https://doi.org/10.1007/s11098-023-02094-3

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Long, R., Butlin, P., Harding, J., Birch, J., Sebo, J., Finlinson, K., Pfau, J., Fish, K., Sims, T., & Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*.

Marcus, G. (2024, January). Where's waldo? the elephant in the room. https://garymarcus.substack.com/p/wheres-waldo-the-elephant-in-the

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175–183.

Millar, A. (2004). *Understanding people: Normativity and rationalizing explanation*. Clarendon Press.

Millière, R. (2022). Adversarial attacks on image generation with made-up words. *arXiv preprint arXiv:2208.04135*.

Millière, R. (2025). Normative conflicts and shallow AI alignment [Forthcoming]. *Philosophical Studies*.

Morrison, J. (2025). *Algorithms for neural networks* [Draft manuscript].

Morton, J. M. (2017). Reasoning under scarcity. *Australasian Journal of Philosophy*, *95*(3), 543–559.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in cognitive sciences*, *25*(9), 757–775.

Musslick, S., & Masís, J. (2023). Pushing the bounds of bounded optimality and rationality. *Cognitive Science*, *47*(4), e13259.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search.

Norton, J. D. (2003). A little survey of induction.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., & Belinkov, Y. (2024). Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*.

Roskies, A. L. (2016). Decision making and self-governing systems. *Neuroethics*, 1–13. https://doi.org/10.1007/s12152-016-9280-9

Schulze, C., Aka, A., Bartels, D. M., Bucher, S. F., Embrey, J. R., Gureckis, T. M., Häubl, G., Ho, M. K., Krajbich, I., Moore, A. K., Oettingen, G., Ongchoco, J. D., Oprea, R., Reinholtz, N., & Newell, B. R. (2025). A timeline of cognitive costs in decision-making. *Trends in Cognitive Sciences*, *29*(9), 827–834. https://doi.org/10.1016/j.tics.2025.04.004

Schurz, G. (2019). *Hume's problem solved: The optimality of meta-induction*. Mit Press.

Schurz, G., & Thorn, P. D. (2016). The revenge of ecological rationality: Strategy-selection by meta-induction within changing environments. *Minds and Machines*, *26*(1), 31–59.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Silverstein, M. (2017). Agency and normative self-governance. *Australasian Journal of Philosophy*, *95*(3), 517–528.

Stiennon, N., et al. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008–3021.

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Sutton, R. (2019). The bitter lesson. *Incomplete Ideas (blog)*, *13*(1), 38.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, *10*(1), 3770.