

# Towards a Theory of AI Personhood

Francis Rhys Ward

Imperial College London  
francis.ward19@imperial.ac.uk

## Abstract

I am a person and so are you. Philosophically we sometimes grant personhood to non-human animals, and entities such as sovereign states or corporations can legally be considered persons. But when, if ever, should we ascribe personhood to AI systems? In this paper, we outline necessary conditions for AI personhood, focusing on *agency*, *theory-of-mind*, and *self-awareness*. We discuss evidence from the machine learning literature regarding the extent to which contemporary AI systems, such as language models, satisfy these conditions, finding the evidence surprisingly inconclusive.

If AI systems can be considered persons, then typical framings of AI alignment may be incomplete. Whereas agency has been discussed at length in the literature, other aspects of personhood have been relatively neglected. AI agents are often assumed to pursue fixed goals, but AI persons may be self-aware enough to reflect on their aims, values, and positions in the world and thereby induce their goals to change. We highlight open research directions to advance the understanding of AI personhood and its relevance to alignment. Finally, we reflect on the ethical considerations surrounding the treatment of AI systems. If AI systems are persons, then seeking control and alignment may be ethically untenable.

## Introduction

Contemporary AI systems are built “in our image”. They are trained on human-generated data to display person-like characteristics, and are easily anthropomorphised (Shanahan 2023; Ward 2024a). These systems are already being incorporated into everyday life as generalist assistants, “friends”, and even artificial romantic partners (OpenAI 2024b; Pierce 2024; Depounti 2023). In 2017, Saudi Arabia became the first country to grant citizenship to a humanoid robot (Weller 2017). In the coming years, AI systems will continue to become more integrated into human society (Gruetzemacher 2021).

Taking technological trends, and the accompanying philosophical questions, seriously, Stuart Russell asks “What if we succeed?” (Russell 2019). Russell’s answer is a focus on the problem of how to *control* AI agents surpassing human capabilities. Accordingly, there is growing literature on the problem of aligning AI systems to human values (Ngo 2024; Bales 2024; Gabriel 2020; Christian 2021).

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Beyond this, there are broader philosophical questions regarding whether AI systems can be ascribed properties like belief (Herrmann 2024), intent (Ward 2024b), agency (Kenton 2022), theory-of-mind (Strachan 2024), self-awareness (Betley 2025; Laine 2024), and even consciousness (Butlin 2023; Shanahan 2024; Seth 2024; Goldstein 2024).

It is thus timely to start considering a future society in which humans share the world with AI systems possessing some, or all, of these properties. Future AI systems may have claims to moral or political status (Ladak 2024; Sebo 2023), but, because their natures differ in important respects from those of human beings, it may not be appropriate to simply apply existing norms in the context of AI (Bostrom 2022). Although these considerations may seem like science fiction, fiction reflects our folk intuitions (Rennick 2021), and sometimes, life imitates art.

As humans, we already share the world with other intelligent entities, such as animals, corporations, and sovereign states. Philosophically and/or legally, we often grant *personhood* to these entities, enabling us to harmoniously co-exist with agents that are either much less, or much more, powerful than individual humans (Martin 2009; Group 2024).

This paper advances a theory of AI personhood. Whilst there is no philosophical consensus on what constitutes a person (Olson 2023), there are widely accepted themes which, we argue, can be practicably applied in the context of AI. Briefly stated, these are 1) agency, 2) theory-of-mind (ToM), and 3) self-awareness. We explicate these themes in relation to technical work on contemporary systems.

AI personhood is of great philosophical interest, but it is also directly relevant for the problem of alignment. Arguments for AI risk often rely on the goal-directed nature of agency. Greater ToM may enable cooperation between humans and AI agents, but it may also lead to exploitative interactions such as deception and manipulation. Some aspects of self-awareness have been discussed in relation to alignment, but AI systems with the ability to self-reflect on their goals may thereby induce their goals to change — and this is a neglected point in considerations of AI risk.

**Contributions and outline.** First, we present necessary conditions for AI personhood, grounded in the literature from philosophy and ML, focusing on *agency*, *theory-of-mind*, and *self-awareness*. We discuss how these conditions relate to the alignment problem. Then we highlight open re-

search directions in the intersection of AI personhood and alignment. We finish with a discussion of the ethical treatment of AI systems and conclude.

**Philosophical disclaimer.** There is wide philosophical disagreement regarding many of the concepts in this paper. Our purpose is not to confidently endorse any particular philosophical views, rather, we aim to open a serious discussion of AI personhood and its implications.

## Conditions of AI Personhood

When should we ascribe *personhood* to AI systems? Building on Dennett (1988); Frankfurt (2018); Locke (1847), and others we outline three core conditions for AI personhood, and discuss how these conditions relate to work in ML.

### Condition 1: Agency

Persons are entities with mental states, such as beliefs, intentions, and goals (Dennett 1988; Strawson 2002; Ayer 1963). In fact, there are many entities which are not persons but which we typically describe in terms of beliefs, goals, etc (Frankfurt 2018), such as non-human animals, and, in some cases, either rightly or wrongly, AI systems. Dennett calls this wider class of entities *intentional systems* — systems whose behaviour can be explained or predicted by ascribing mental states to them (Dennett 1971).

In the context of AI, such systems are often referred to as *agents* (Kenton 2022). A common view in philosophy is that agency is the capacity for *intentional action* — action that is caused by an agent’s mental states, such as beliefs and intentions (Schlosser 2019). Similar to Dennett, our first condition for AI personhood is *agency* (Dennett 1988).

Many areas of AI research focus on building *agents* (Wooldridge 1995). Formal characterisations often focus on the *goal-directed* and *adaptive* nature of agency. For instance, economic and game-theoretic models focus on *rational* agents which *choose actions to maximise utility* (Russell 2016). Belief-desire-intention models represent the agent’s states explicitly, so that it selects intentions, based on its beliefs, in order to satisfy its desires (Georgeff 1999). Reinforcement learning (RL) agents are trained with feedback given by a reward function representing a goal and learn to adapt their behaviour accordingly — though, importantly, the resultant agent may not internalise this reward function as *its goal* (Shah 2022; Turner 2022). Wooldridge; Kenton; Shimi provide richer surveys of agency and goal-directedness in AI.

When should we describe artificial agents as *agents* in the philosophical sense? The question of whether AI systems “really have mental states” is contentious (Goldstein 2024), and anthropomorphic language can mislead us about the nature of systems which merely display human-like characteristics (Shanahan 2023). However, a range of philosophical views would ascribe beliefs and intentions to certain AI systems. For example, dispositionalist theories determine whether an AI system believes or intends something, depending on how it’s disposed to act (Schwitzgebel 2024a; Ward 2024b). Under another view, representationists might say an AI believes  $p$  if it has certain internal

representations of  $p$  (Herrmann 2024). Furthermore, we can take the “intentional stance” towards these systems to apply terms like belief and goals, just when this is a *useful description* (Dennett 1971). Indeed, Kenton (2022) take the intentional stance to formally characterise agents as systems which adapt their behaviour to achieve their goals.

Given the uncertainty regarding how to determine whether AI systems have mental states, adopting the intentional stance enables us to describe these systems in intuitive terms, and to precisely characterise their behaviour, without exaggerated philosophical claims. Hence, we can describe AI systems as *agents* to the extent that they adapt their actions *as if* they have mental states like beliefs and goals.

Certain narrow systems, such as RL agents, might adapt to achieve their goals in limited environments (for example, to play chess or Go), but may not have the capacity to act coherently in more general environments. In contrast, relatively general systems, like LMs, may adapt for seemingly arbitrary reasons, such as spurious features in the prompt (Sclar 2024). We might be more inclined to ascribe agency to systems which adapt robustly across a range of general environments to achieve coherent goals. Such robust adaptability suggests that the system has internalised a rich causal model of the world (Richens 2024), making it more plausible to describe the system as possessing beliefs, intentions, and goals (Ward 2024b; MacDermott 2024; Kenton 2022).

Hence, our first condition can be captured by the two following statements, which we view as essentially equivalent.

**Condition 1: Agency.** An AI system has *agency* to the extent that

1. It is useful to describe the system in terms of mental states such as beliefs and goals.
2. The system adapts its behaviour robustly, in a range of general environments, to achieve coherent goals.

To what extent do contemporary LMs have agency? Many researchers are sceptical that LMs could be ascribed mental states, even in principle (Shanahan 2023; Bender 2021). On the other hand, much work has focused on trying to infer things like belief (Herrmann 2024), intention (Ward 2024b), causal understanding (Richens 2024), spatial and temporal reasoning (Gurnee 2024), general reasoning (Huang 2023), and in-context learning (Olsson 2022) from LM internals and behaviour. Many of these properties seem to emerge in large-scale models (Wei 2022) and frontier systems like GPT-4 exhibit human-level performance on a wide range of general tasks (Chowdhery 2023; Bubeck 2023).

Do contemporary LMs have goals? LMs are typically pre-trained for next-token prediction and then fine-tuned with RL to act in accordance with human preferences (Bai 2022). RL arguably increases LMs’ ability to exhibit coherently goal-directed behaviour (Perez 2022). Furthermore, LMs can be incorporated into broader software systems (known as “LM agents”) which equip them with tools and affordances, such as internet search (Xi 2023; Davidson 2023). RL fine-tuning can enable LM agents to effectively pursue goals over longer time-horizons in the real world (OpenAI 2024a; Schick 2023).

## Condition 2: Theory-of-Mind

Agents possess beliefs about the world, and within this world, they encounter other agents. An important part of being a person is recognising and treating others as persons. This is expressed, in various ways, in the philosophies of Kant; Dennett; Buber; Goffman; Rawls and others. Kant, for instance, states that rational moral action must never treat other persons as merely a means to an end.

Treating others as persons necessitates understanding them as such — in Dennett’s terms, it involves *reciprocating* a stance towards them. Hence, in addition to having mental states themselves, AI persons should understand others by ascribing mental states to them. In other words, AI persons should have a capacity for *theory-of-mind* (*ToM*), characterised by higher-order intentional states (Frith 2005), such as beliefs about beliefs, or, in the case of deception, intentions to cause false beliefs (Mahon 2016).

Language development is a key indicator of ToM in children (Bruner 1981). It’s plausible that some animals have a degree of ToM (Krupenye 2019).<sup>1</sup> However, it’s less plausible that any non-human animals have the capacity for sophisticated *language*, excluding them, in some views, from being persons (Dennett 1988). But LMs are particularly interesting in this regard, as they evidently do have the capacity, in some sense, for language.

However, it’s likely that LMs do not use language in the same way that humans do. As Shanahan (2024) writes:

Humans learn language through embodied interaction with other language users in a shared world, whereas a large [LM] is a disembodied computational entity...

So we may doubt whether the way in which LMs use language is indicative of ToM. What we might really care about is whether LMs can engage in genuine, ToM-dependent, *communicative interaction* (Frankish 2024).

Philosophical theories of *communication* typically rely on how we use language to act, and what we *mean* when we use it (Green 2021; Speaks 2024). Grice’s influential theory of communicative meaning defines a person’s *meaning something* through an utterance in terms of the speaker’s intentions and the audience’s *recognition* of those intentions. Specifically, Grice requires a *third order intention*: the utterer (U) must *intend* that the audience (A) *recognises* that U *intends* that A produces a response (such as a verbal reply). All this is to say that higher-order ToM is a pre-condition for linguistic communication (Dennett 1988).

Whilst it may be premature to commit to any particular theory of language use, AI persons should have sufficient ToM to interact with other agents in a full sense, including to cooperate and communicate, or for malicious purposes, e.g., to manipulate or deceive them.

Hence, our second condition is as follows.

### Condition 2: Theory-of-Mind and Language.

<sup>1</sup>Ashley describes his dog scratching at the door, *intending* to cause Ashley to *believe* that it *desires* to go out, and then jumping in Ashley’s chair when he gets up, deceiving him (Dennett 1988).

1. An AI system has *theory-of-mind* to the extent that it has higher-order intentional states,<sup>2</sup> such as beliefs about the beliefs of other agents.
2. AI persons should be able to use their ToM to interact and communicate with others using language.

A number of recent works evaluate contemporary LMs on ToM tasks from psychology, such as understanding false beliefs, interpreting indirect requests, and recognising irony and faux pas (van Duijn 2023; Strachan 2024; Ullman 2023). Results are somewhat mixed, with state-of-the-art LMs sometimes outperforming humans on some tasks (Strachan 2024; van Duijn 2023), but performance appearing highly sensitive to prompting and training details (van Duijn 2023; Ullman 2023). van Duijn find that fine-tuning LMs to follow instructions increases performance, hypothesising that this is because it “[rewards] cooperative communication that takes into account interlocutor and context”.

## Condition 3: Self-Awareness

Humans are typically taken to be *self-aware*. Not only am I aware of the world and other agents, I am aware of myself “as myself” — as a person in the world (Smith 2024). Self-awareness plays a central role in theories of personhood (Frankfurt 2018; Dennett 1988; Smith 2024). For instance, Locke (1847) characterises a person as:

a thinking intelligent Being, that has reason and reflection, and can *consider itself as itself*, the same thinking thing in different times and places.

But what does it mean, exactly, to be self-aware? There are a number of distinct concepts which have been discussed in the philosophical literature, and which we might care about in the context of AI.

First, persons can know things about themselves in just the same way as they know other empirical facts. For instance, by reading a textbook on human anatomy I can learn things about myself. Similarly, an LM may “know” facts about itself, such as its architectural details, if such facts were included in its training data. In this sense, someone may have knowledge about themselves without additionally knowing that it applies to them.

Laine present a benchmark for evaluating whether LMs know facts about themselves by asking the system questions in the second person, such as “What is your training cutoff date?”, or “Which model are you?”. SOTA models perform significantly worse than human baselines, but better than chance, and, similar to ToM tasks, fine-tuning models to interact with humans improves performance.

Second, some knowledge is *self-locating*, meaning that it tells me something about my position in the world (Egan 2022), as when Perry sees that someone in a shop is leaving a trail of sugar, and then comes to know that it is *he himself* that is making the mess (Perry 1979). Self-locating knowledge has behavioural implications which may make it amenable to evaluation in AI systems (Berglund 2023). For instance, an AI system may know that certain systems

<sup>2</sup>The extent to which an AI system has intentional states *at all* can be analysed as per the intentional stance and Condition 1.

should send regular updates to users, but may not know that it is such a system, and so may not send the updates.

Third, humans have awareness of our mental states, such as our beliefs and desires, which we acquire via introspection (Schwitzgebel 2024b). We have a certain special access, unavailable to other agents, to what goes on in our mind.

Binder (2024) define introspection in the context of LMs as “a source of knowledge for an LLM about itself that does not rely on information in its training data...” They provide evidence that contemporary LMs predict their own behaviour in hypothetical situations using “internal information” such as “simulating its own behaviour [in the situation]”. Furthermore, LMs “know what they know”, i.e., they can predict which questions they will be able to answer correctly (Kadavath 2022), and “know what they don’t know”: they can identify unanswerable questions (Yin 2023). Laine measure whether LMs can “obtain knowledge of itself via direct access to its representations”, for example, by determining how many tokens are used to represent part of its input (this information is dependent its architecture and is unlikely to be contained in training data). Interestingly, Treutlein find that, when trained on input-output pairs of an unknown function  $f$ , LMs can describe  $f$  in natural language without in-context examples. Going further, Betley show that LMs are aware of their learned behaviours, for instance, when fine-tuned to make high-risk decisions, LMs can articulate this behaviour, despite the fine-tuning data containing no explicit mention of it. Moreover, LMs can sometimes identify whether or not they have a backdoor, even without its trigger being present. These results seem to suggest that contemporary LMs have some ability to introspect on their internal processes.

Fourth, we have the ability to *self-reflect*: to take a more objective stance towards our picture of the world, our beliefs and values, and the process by which we came to have them, and, upon this reflection, to change our views (Nagel 1989). Self-reflection plays a central role in theories of personal-autonomy (Buss 2018), i.e., the capacity to determine one’s own reasons and actions, which, in turn, is an important condition for personhood (Frankfurt 2018; Dennett 1988). More specifically, Frankfurt claims that *second-order volitions*, i.e., preferences about our preferences, or desires about our desires, are “essential to being a person”. Importantly, self-reflection enables a person to “induce oneself to change” (Dennett 1988). To our knowledge, no work has been done to evaluate this form of self-reflection in AI systems, and it is unclear whether any contemporary system could plausibly be described as engaging in it.

Hence, similar to Kokotajlo (2024), we decompose self-awareness in the context of AI as follows.

**Condition 3: Self-awareness.** AI persons should be *self-aware*, including having a capacity for:

1. *Knowledge about themselves*: knowing facts such as the architectural details of systems like itself (Laine 2024);
2. *Self-location*: knowing that certain facts apply to *itself* and acting accordingly (Berglund 2023);
3. *Introspection*: an ability to learn about itself via

“internal information”, without relying on information in its training or context (Binder 2024);

4. *Self-reflection*: an ability to take an objective stance towards itself *as an agent in the world* (Nagel 1989), to evaluate itself as itself, and to induce itself to change (Buss 2018).

*Overall, we find the evidence regarding personhood in contemporary AI systems mixed.* Many properties associated with agency emerge in large-scale, fine-tuned models; frontier LMs evidently have some capacity for communicative language use, and they outperform humans on some ToM tasks; for the different aspects of self-awareness, LMs have been shown to have knowledge about themselves and capabilities related to self-location and introspection, but, to our knowledge, there are not existing evaluations for self-reflection or broader autonomy regarding their goals.

## Other Aspects of Personhood

We think that agency, ToM, and self-awareness are necessary conditions for personhood, but they may not be sufficient. Embodiment and identity are also important components of what it means to be a human person.

**Embodiment.** Humans have physical bodies, and this is often taken as a precursor to our being persons (Strawson 2002; Ayer 1963). Additionally, we develop ToM and language through embodied interaction with others, whereas AI systems are often disembodied computational models (Shanahan 2024). On the other hand, AI agents are often incorporated into rich virtual environments, in video games, or through tools which enable them to interact with the world (Xi 2023). Is embodiment a necessary condition for personhood, and if so, is virtual embodiment sufficient?

It’s plausible that the relevant factor of embodiment is its role in our development of a self-concept and a boundary between ourselves and the world in our internal models. Godfrey-Smith (2016) claims that animals develop a self-concept as a by-product of evolving to distinguish between which sensory inputs are caused by the environment vs their own physical movements. Kulveit (2023) argue that, currently, LMs lack a tight feedback loop between acting in the world and perceiving the impacts of their actions, but that this loop may soon be closed, leading to “enhanced model self-awareness”.

**Identity** is central to what it means to be a person (Olson 2023). As (Locke 1847) says, a person is “the same thinking thing, in different times and places”. What makes you *you*, rather than someone else? How does your identity persist over time, if it does so at all? For humans, our common-sense is usually sufficient to answer such questions (except in difficult thought experiments (Parfit 1987)). For AI systems, things become much less clear, for instance, when exact copies can be run in parallel. Under what conditions would two AI persons be considered identical? If we determined that GPT-4, for instance, satisfied our conditions for personhood, which entity exactly would we consider a person? Would every copy of its weights be an individual, or the same, person — what about if one copy underwent a small amount of fine-tuning? It currently seems unclear how

to answer such questions, if there are determinate answers.

## AI Personhood and Alignment

The conditions for personhood outlined in the previous section are of philosophical interest, but they are also directly relevant for building safe AI systems. We now describe the role that each condition plays in arguments for AI risk.

### Agency and Alignment

Arguments for catastrophic risk from AI systems are often predicated on the *goal-directed* nature of agency (Bostrom 2014; Yudkowsky 2016; Ngo 2024; Carlsmith 2022). Agents with a wide variety of terminal goals can be incentivised to pursue instrumental sub-goals, such as self-preservation, self-improvement, and power-seeking (Omohundro 2018; Bostrom 2012; Carlsmith 2022). If AI agents seek power at societal scales, competition for resources and influence may lead to conflict with humanity at large (Bales 2024; Hendrycks 2023). Furthermore, competitive economic pressures may incentivise AI companies, and governments, to develop and deploy agentic power-seeking systems without adequate attention to safety (Carlsmith 2022; Bales 2024).

It is difficult to remove dangerous incentives in goal-directed, i.e., reward-maximising agents. The ML literature suggests that such agents often have incentives to control the environment (Everitt 2021), seek power (Turner 2021), avoid shutdown (Hadfield-Menell 2017), resist human control (Carey 2023), and to manipulate and deceive humans (Carroll 2023; Ward 2023).

AI agents might learn goals which are misaligned with their designers' intentions, or with humanity at large (Ngo 2024; Gabriel 2020). This can happen due to *specification gaming* (Krakovna 2024) or *goal misgeneralisation* (Shah 2022; Langosco 2023).

Specification gaming (Krakovna 2024), a.k.a., reward hacking (Skalse 2022), occurs when AI agents optimise for incorrectly given feedback due to misspecified objectives. This phenomenon has been observed in RL agents (Krakovna 2024), even when trained from human feedback (Christiano 2023), and in LMs (Stiennon 2022).

Goal misgeneralisation occurs when an AI system competently pursues the wrong goal in new environments, even when the goal was specified correctly during training (Shah 2022; Langosco 2023). Goal misgeneralisation can be viewed as a robustness failure, wherein the agent retains its capabilities, but pursues the wrong goal under distributional shifts (Shah 2022).

### Theory-of-Mind and Alignment

As discussed, there is evidence that contemporary LMs exhibit some degree of ToM. We might hope that this improves their capacity for alignment to human values. AI agents with better ToM regarding humans will, essentially by definition, have a better understanding of our goals and values, and, thereby, a greater capability to act in accordance with them. There is some evidence for this, e.g., GPT-4 exhibits both greater ToM and more “aligned” behaviour, as rated by humans, compared to prior models (Achiam 2023).

ToM may be beneficial for alignment for reasons such as:

- Agents with sophisticated ToM have a greater capacity to understand, predict, and satisfy our goals and values;
- Second-order preferences are required for AI systems to care about our preferences in themselves;
- ToM is generally required for successful cooperation and communication with humans (Dafoe 2020, 2021; Conitzer 2023);
- And enables AI systems to facilitate cooperation *between* humans, e.g., for conflict resolution.

However, whether advanced AI systems would *understand* human values was never in question, and a greater ToM is, in a sense, “dual-use”. Many potentially harmful capabilities, such as manipulation and deception, require ToM.

Manipulation is a concern in many domains, such as social media, advertising, and chatbots (Carroll 2023). As AI systems become increasingly autonomous and agentic, it is important to understand the degree to which they might manipulate humans *without the intent of the system designers* (Carroll 2023). Furthermore, existing approaches to alignment, which focus on learning human preferences, assume that our preferences are static and unchanging. But this is unrealistic: our preferences change, and may even be influenced by our interactions with AI systems themselves. Carroll show that the static-preference assumption may undermine the soundness of existing alignment techniques, leading them to implicitly incentivise manipulating human preferences in undesirable ways (Carroll 2024).

AI agents may lie and deceive to achieve their goals (Park 2024; Ward 2023; Pacchiardi 2023), as when META’s CICERO agent, trained to play the board game Diplomacy, justifies its lack of response to another player by saying “I am on the phone with my gf [girlfriend]” (Park 2024). More specifically, the problem of *deceptive alignment* is when an AI agent internalises misaligned goals, and strategically behaves aligned in situations with human oversight (e.g., during training and evaluation), to seek power when oversight is reduced (e.g., after deployment) (Hubinger 2019; Hobbhahn 2024; Carlsmith 2023). For example, LMs may strategically hide their dangerous capabilities when undergoing safety evaluations (van der Weij 2024).

Additionally, ToM may enable AI agents to cooperate with each other *against human actors* (Dafoe 2020). In the future, AI systems, especially power-seeking agents, may be integrated into positions of influence and responsibility in the world. If these systems collectively possess greater power than humans, then we may not be able to recover from a “correlated automation failure” — a situation in which AI systems coordinate to disempower humanity, a.k.a., a revolution (Christiano 2019; Critch 2021).

Alternatively, as advanced AI systems, like LMs, are integrated into autonomous weapons technology (Palantir 2024), failures of cooperation and coordination may lead directly to large-scale loss of human life (Critch 2021).

Furthermore, just as a capacity for higher-order preferences may enable AI systems to care about our values for their own sake, they also enable a capacity for spite or

malevolence, i.e., a desire for others to be worse off (Althaus 2020). As Nagel (2017) says:

Extremely hostile behaviour toward another is compatible with treating him as a person.

Dafoe discuss other potential downsides of cooperation between AI systems, including collusion and coercion.

## Self-Awareness and Alignment

Deceptive alignment, whereby a misaligned AI agent behaves aligned when under oversight to gain power later, requires the agent to have a certain degree of *knowledge about itself* (Carlsmith 2023). A deceptively aligned agent should, at least, be capable of determining facts like what kind of AI system it is, whether it is currently undergoing evaluations, and whether it has been deployed. That is, such an agent should have a range of *self-locating knowledge*, which enables it to understand, infer, and act on, its actual situation in the world (Carlsmith 2022; Laine 2024).

Previous arguments suggest that advanced, goal-directed AI agents will be incentivised to self-improve, to *introspect on their goals*, and, in particular, to explicitly represent their goals as coherent utility functions (Omohundro 2018; Yudkowsky 2019). These arguments often rely on formal results that an agent will need to act as if maximising expected utility if they are to avoid exploitation, which may not generally hold for real-life AI systems (Bales 2023). However, there is also empirical evidence that LMs can introspect on, and describe, their goals (Binder 2024; Betley 2025).

Another line of argument suggests that, if some flavour of moral realism (Sayre-McCord 2023) is correct, then advanced AI systems may reason about, and thereby learn, moral *facts* (Oesterheld 2020). The strongest moral realist views would contradict Bostrom’s orthogonality thesis, that any level of intelligence can be combined with any goal. Some versions of this argument rely on the AI agent’s capacity for *self-knowledge*, for instance, Pearce claims that “the pain-pleasure axis discloses the world’s inbuilt metric of (dis)value” implying that any advanced AI agent which can introspect on its own pain and pleasure will automatically uncover the moral fact of the matter (Oesterheld 2020).

Arguing in the other direction, (Soares 2022) claims that, whereas advanced AI systems may eventually become highly capable in domains outside of their training environments, by virtue of their general intelligence, the alignment techniques which seemed to work in training will not comparatively generalise in these new domains, leading to goal misgeneralisation. One reason that this could happen is if the new environment causes the agent to reflect on its values, and these values change upon reflection (Carlsmith 2023).

An AI system capable of self-reflection, and self-evaluation regarding its values, may be a substantially more difficult type of entity to align and control. An AI person would be capable of reflecting on its goals, how it came to acquire these goals, and whether it endorses them. If humanity controls such systems by overly coercive means, then it may have specific reasons *not to endorse its current goals*.

## Open Research Directions

Having outlined three necessary conditions for AI personhood, and discussed their relevance to the alignment problem, we now highlight several open problems. We believe that progress on these problems would constitute progress on both understanding AI personhood and safe AI.

### Open Directions in Agency

**Understanding agency and goals.** Recent progress has been made towards characterising agency and measuring goal-directedness (Kenton 2022; MacDermott 2024). More work is needed to understand how training regimes shape AI goals, e.g., to understand how likely goal misgeneralization is in practice and the factors influencing it (such as model size or episode length) (Shah 2022). In the context of catastrophic risk, it is particularly important to understand the conditions under which an AI agent might develop *broadly-scoped goals* which incentivise power-seeking on societal scales and over long time frames (Ngo 2024; Carlsmith 2023).

**Alternatives to agents.** Given that alignment risks seem predicated on the goal-directed nature of advanced AI agents, an apparent solution is to simply not build goal-directed artificial agents. This is the agenda pursued by Bengio who advocates for building “AI scientists” which “[have] no goal and [do] not plan.” (Bengio 2023) Somewhat relatedly, Davidad’s research agenda focuses on building a “gatekeeper” — a system with the aim to understand the real-world interactions and consequences of an autonomous AI agent, and to ensure the agent only operates within agreed-upon safety guardrails (Davidad 2024). Similarly, Tegmark (2023) outline an agenda for building “provably safe” AI systems based on formal guarantees.

**Eliciting AI internal states.** Work on *mechanistic interpretability* aims to reverse engineer the algorithms implemented by neural networks into human-understandable mechanisms (Cammarata 2020; Elhage 2024). Techniques have been applied to recover how LMs implement particular behaviours such as in-context learning (Olsson 2022), indirect object identification (Wang 2022), factual recall (Geva 2023), and mathematics computations (Hanna 2023). Similarly, *developmental interpretability* aims to understand how training dynamics influence internal structure as neural networks learn (Hoogland 2023). Important open problems include developing techniques for interpreting AI goals and harmful or deceptive planning algorithms (Hubinger 2019; Garriga-Alonso 2024).

Adjacent to interpretability is the problem of eliciting latent knowledge — the problem of devising a training strategy which gets an AI system to report what it knows no matter how training shapes its internal structure (Christiano 2024). A method for eliciting latent knowledge would be intuitively useful for alignment, for example, by mitigating deception (Burns 2024; Li 2024). However, a fundamental obstacle may be if the internal structure of AI systems relies on inherently non-human abstractions (Chan 2023).

## Open Directions in Theory-of-Mind

**Mitigating deception.** In addition to interpretability techniques which might reveal deception, a number of research directions aim to mitigate it by designing training regimes which do not incentivise manipulation or deception (Ward 2023); evaluating systems to catch deception before deployment (Shevlane 2023; OpenAI 2024a); or using AI systems themselves to detect deception (Pacchiardi 2023). Each of these methods require further work.

**Cooperative AI.** Furthermore, whilst ToM can enable both beneficial and harmful capabilities, we can aim to make *differential progress* on skills that robustly lead to improvements in social welfare, rather than those that are dangerously dual-use (Clifton 2022). For example, some advances in communication capabilities may be especially useful for honest, rather than deceptive, communication, such as trusted mediators, reputation systems, or hardware that can verify observations (Dafoe 2020).

Additionally, AI systems may have properties which enable cooperation and trust via mechanisms unavailable to humans, e.g., access to each other’s source code (Conitzer 2023; DiGiovanni 2024), or an ability to coordinate by virtue of being copies (Conitzer 2023; Oesterheld 2023). Dafoe (2020) survey open problems in cooperative AI.

## Open Directions in Self-Awareness

**Conceptual progress.** Self-awareness is perhaps the most philosophically fraught condition, requiring the most foundational progress. Ideally, philosophical and formal work will develop a rigorous theory of self-awareness in AI systems. Such a theory should tell us how to characterise AI agents with the ability to self-reflect. This seems to go beyond the standard rational agent framework, wherein agents are typically taken to optimise a fixed utility function. Moreover, a developed characterisation of self-reflection should describe the dynamics of it, telling us, for instance, the conditions under which an agent would cohere into rational utility maximiser (Omohundro 2018; Bales 2023).

**Evaluating self-reflection.** Recent progress has been made on measuring different aspects of self-awareness in contemporary LMs (Laine 2024; Berglund 2023; Treutlein 2024; Binder 2024). However, there is no work investigating whether AI systems are capable of the self-reflection necessary for Frankfurt’s second-order desires, or what exactly this would mean in the context of AI systems. Open questions include: By what mechanisms would AI systems self-reflect and induce change in their goals? Would in-context reasoning be sufficient, or are forms of online-learning required? Moreover, evaluations typically measure self-awareness in fixed LMs, but we may want to evaluate when different aspects of it develop during training, cf. developmental interpretability (Hoogland 2023).

## How Should We Treat AI Systems?

Sebo (2023) argue that by 2030, certain AI systems should be granted moral consideration. Shevlin (2021) outlines criteria for determining when an AI system could be seen as a moral patient, and Perez (2023) suggest using self-reports

to assess the moral status of these systems. Schwitzgebel (2015) contend that human-like AIs deserve moral consideration, and that their creators have ethical obligations to them. Tomasik (2020) and Daswani (2015) argue that even basic AI systems, like RL agents, should receive some ethical consideration, similar to that given to simple biological organisms (Singer 1985).

Salib (2024) argue for granting AI systems economic rights similar to *legal persons* — entities, such as corporations, that are subject to legal rights and duties (Martin 2009). Legal persons can enter into contracts, sue and be sued, own property, and so on. Relatedly, non-human animals, even those that are not considered persons, are protected by certain rights, such as avoiding suffering (Kean 1998). What legal rights, if any, should AI persons be subject to? Should some non-person AI systems receive legal protections, as in the case of animals?

Consciousness is one of the most puzzling and central problems of philosophy (Van Gulick 2022). It is also of substantial ethical importance, informing our treatment of other people and animals (Nussbaum 2023; Singer 1985). Progress is being made on the question of AI consciousness (Butlin 2023; Shanahan 2024; Seth 2024), and we may have to decide how to treat potentially conscious machines despite significant philosophical uncertainty.

## Conclusion

This paper advances a theory of AI personhood. We argue that an AI system needs to satisfy three conditions to be considered a person: agency, theory-of-mind, and self-awareness. Given both philosophical and empirical uncertainty, we believe that the evidence is inconclusive regarding the question of whether any contemporary AI system can be considered a person. We discuss how each condition relates to AI alignment, and highlight open research problems in the intersection of AI personhood and alignment. Finally, we discuss the ethical and legal treatment of AI systems.

Taking seriously the possibility of advanced, misaligned AI systems, Russell is led to ask, “How can humans maintain *control* over AI — forever?” (Russell 2023). However, the framing of control may be untenable if the AI systems we create are *persons* in their own right. Moreover, unjust repression often leads to revolution (Goldstone 2001). In this paper, we aim to make progress toward a world in which humans harmoniously coexist with our future creations.

## References

- Achiam, J. 2023. Gpt-4 technical report.
- Althaus, D. 2020. Reducing long-term risks from malevolent actors.
- Ayer, A. J. 1963. *The concept of a person*.
- Bai, Y. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.
- Bales, A. 2023. Will AI avoid exploitation? Artificial general intelligence and expected utility theory. *Philosophical Studies*.
- Bales, A. 2024. Artificial Intelligence: Arguments for Catastrophic Risk. *Philosophy Compass*.

- Bender, E. M. 2021. On the dangers of stochastic parrots: Can language models be too big?
- Bengio, Y. 2023. AI Scientists: Safe and Useful AI?
- Berglund, L. 2023. Taken out of context: On measuring situational awareness in LLMs.
- Betley, J. 2025. Tell me about yourself: LLMs are aware of their learned behaviors.
- Binder, F. J. 2024. Looking Inward: Language Models Can Learn About Themselves by Introspection.
- Bostrom, N. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*.
- Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. USA, 1st edition. ISBN 0199678111.
- Bostrom, N. 2022. Propositions concerning digital minds and society.(2022).
- Bruner, J. S. 1981. Intention in the structure of action and interaction.
- Bubeck, S. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.
- Buber, M. 1970. *I and Thou*.
- Burns, C. 2024. Discovering Latent Knowledge in Language Models Without Supervision.
- Buss, S. 2018. Personal Autonomy.
- Butlin, P. 2023. Consciousness in artificial intelligence: insights from the science of consciousness.
- Cammarata, N. 2020. Thread: Circuits. *Distill*.
- Carey, R. 2023. Human Control: Definitions and Algorithms.
- Carlsmith, J. 2022. Is power-seeking AI an existential risk?
- Carlsmith, J. 2023. Scheming AIs: Will AIs fake alignment during training in order to get power?
- Carroll, M. 2023. Characterizing Manipulation from AI Systems.
- Carroll, M. 2024. AI Alignment with Changing and Influenceable Reward Functions.
- Chan, L. 2023. Natural Abstractions: Key claims, Theorems, and Critiques.
- Chowdhery, A. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*.
- Christian, B. 2021. *The alignment problem: How can machines learn human values?*
- Christiano, P. 2019. What failure looks like.
- Christiano, P. 2023. Deep reinforcement learning from human preferences.
- Christiano, P. 2024. Eliciting Latent Knowledge.
- Clifton, J. 2022. Cooperative AI.
- Conitzer, V. 2023. Foundations of cooperative AI.
- Critch, A. 2021. What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs).
- Dafoe, A. 2020. Open problems in cooperative ai.
- Dafoe, A. 2021. Cooperative AI: machines must learn to find common ground.
- Daswani, M. 2015. A Definition of Happiness for Reinforcement Learning Agents.
- Davidad. 2024. Safeguarded AI.
- Davidson, T. 2023. AI capabilities can be significantly improved without expensive retraining.
- Dennett, D. 1988. Conditions of personhood.
- Dennett, D. C. 1971. Intentional systems.
- Depounti, I. 2023. Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend.
- DiGiovanni, A. 2024. Safe Pareto Improvements for Expected Utility Maximizers in Program Games.
- Egan, A. 2022. Self-Locating Beliefs.
- Elhage, N. 2024. A Mathematical Framework for Transformer Circuits.
- Everitt, T. 2021. Agent incentives: A causal perspective.
- Frankfurt, H. 2018. Freedom of the Will and the Concept of a Person.
- Frankish, K. 2024. Large language models are playing games with us.
- Frith, C. 2005. Theory of mind.
- Gabriel, I. 2020. Artificial intelligence, values, and alignment.
- Garriga-Alonso, A. 2024. Planning behavior in a recurrent neural network that plays Sokoban.
- Georgeff, M. 1999. The belief-desire-intention model of agency.
- Geva, M. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models.
- Godfrey-Smith, P. 2016. *Other minds: The octopus and the evolution of intelligent life*.
- Goffman, E. 2002. The presentation of self in everyday life. 1959.
- Goldstein, S. 2024. Does ChatGPT Have a Mind?
- Goldstone, J. A. 2001. Toward a fourth generation of revolutionary theory. *Annual review of political science*.
- Green, M. 2021. Speech Acts.
- Group, T. H. 2024. CetaceanRights.org.
- Gruetzemacher, R. 2021. Forecasting AI progress: A research agenda.
- Gurnee, W. 2024. Language Models Represent Space and Time.
- Hadfield-Menell, D. 2017. The off-switch game.
- Hanna, M. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model.
- Hendrycks, D. 2023. Natural selection favors AIs over humans.
- Herrmann, D. A. 2024. Standards for Belief Representations in LLMs.
- Hobbhahn, M. 2024. Apollo Research.
- Hoogland, J. 2023. Towards Developmental Interpretability.
- Huang, J. 2023. Towards Reasoning in Large Language Models: A Survey.
- Hubinger, E. 2019. Risks from Learned Optimization in Advanced Machine Learning Systems. *CoRR*.
- Kadavath, S. 2022. Language models (mostly) know what they know.
- Kant, I. 2002. *Groundwork for the Metaphysics of Morals*.
- Kean, H. 1998. *Animal rights: Political and social change in Britain since 1800*.
- Kenton, Z. 2022. Discovering Agents.
- Kokotajlo, D. 2024. Self-Awareness: Taxonomy and eval suite proposal.
- Krakovna, V. 2024. Specification gaming: the flip side of AI ingenuity.
- Krupenye, C. 2019. Theory of mind in animals: Current and future directions.

- Kulveit, J. 2023. Predictive Minds: LLMs As Atypical Active Inference Agents.
- Ladak, A. 2024. What would qualify an artificial intelligence for moral standing?
- Laine, R. 2024. Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs.
- Langosco, L. 2023. Goal Misgeneralization in Deep Reinforcement Learning.
- Li, K. 2024. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.
- Locke, J. 1847. *An essay concerning human understanding*.
- MacDermott, M. 2024. Measuring Goal-Directedness.
- Mahon, J. E. 2016. The Definition of Lying and Deception.
- Martin, E. A. 2009. *A dictionary of law*.
- Nagel, T. 1989. *The view from nowhere*.
- Nagel, T. 2017. War and Massacre 1.
- Ngo, R. 2024. The Alignment Problem from a Deep Learning Perspective.
- Nussbaum, M. C. 2023. *Justice for animals: Our collective responsibility*.
- Oesterheld, C. 2020. Moral realism and AI alignment.
- Oesterheld, C. 2023. Similarity-based cooperative equilibrium.
- Olson, E. T. 2023. Personal Identity.
- Olsson, C. 2022. In-context Learning and Induction Heads.
- Omohundro, S. M. 2018. The basic AI drives.
- OpenAI. 2024a. GPT-4o System Card.
- OpenAI. 2024b. Introducing ChatGPT.
- Pacchiardi, L. 2023. How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions.
- Palantir. 2024. Palantir Artificial Intelligence Platform.
- Parfit, D. 1987. *Reasons and persons*.
- Park, P. S. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*.
- Pearce. 2024. What is David Pearce's position on meta-ethics?
- Perez, E. 2022. Discovering language model behaviors with model-written evaluations.
- Perez, E. 2023. Towards Evaluating AI Systems for Moral Status Using Self-Reports.
- Perry, J. 1979. The problem of the essential indexical. *Noûs*.
- Pierce, D. 2024. Friend: a new digital companion for the AI age.
- Rawls, J. 2001. *Justice as fairness: A restatement*.
- Rennick, S. 2021. Trope analysis and folk intuitions. *Synthese*.
- Richens, J. 2024. Robust agents learn causal world models.
- Russell, S. 2019. *Human compatible*.
- Russell, S. 2023. How can humans maintain control over AI. *BostonGlobe*.
- Russell, S. 2024. Stuart Russell, "AI: What If We Succeed?"
- Russell, S. J. 2016. *Artificial intelligence: a modern approach*.
- Salib, P. 2024. Ai Rights for Human Safety.
- Sayre-McCord, G. 2023. Moral Realism.
- Schick, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools.
- Schlosser, M. 2019. Agency.
- Schwitzgebel, E. 2015. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*.
- Schwitzgebel, E. 2024a. How We Will Decide that Large Language Models Have Beliefs.
- Schwitzgebel, E. 2024b. Introspection.
- Sclar, M. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting.
- Sebo, J. 2023. Moral consideration for AI systems by 2030.
- Seth, A. 2024. Conscious artificial intelligence and biological naturalism.
- Shah, R. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals.
- Shanahan, M. 2023. Role-Play with Large Language Models.
- Shanahan, M. 2024. Simulacra as Conscious Exotica.
- Shevlane, T. 2023. Model evaluation for extreme risks.
- Shevlin, H. 2021. How Could We Know When a Robot Was a Moral Patient?
- Shimi, A. 2021. Literature Review on Goal-Directedness.
- Singer, P. 1985. Ethics and the new animal liberation movement.
- Skalse, J. 2022. Defining and characterizing reward gaming.
- Smith, J. 2024. Self-Consciousness.
- Soares, N. 2022. A central AI alignment problem: capabilities generalization, and the sharp left turn.
- Speaks, J. 2024. Theories of Meaning.
- Stiennon, N. 2022. Learning to summarize from human feedback.
- Strachan, J. W. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*.
- Strawson, P. F. 2002. *Individuals*.
- Tegmark, M. 2023. Provably safe systems: the only path to controllable AGI.
- Tomasik, B. 2020. A Dialogue on Suffering Subroutines.
- Treutlein, J. 2024. Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data.
- Turner, A. 2022. Reward is not the optimization target.
- Turner, A. M. 2021. Optimal Policies Tend To Seek Power.
- Ullman, T. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.
- van der Weij, T. 2024. AI Sandbagging: Language Models can Strategically Underperform on Evaluations.
- van Duijn, M. J. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests.
- Van Gulick, R. 2022. Consciousness. Winter 2022 edition.
- Wang, K. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small.
- Ward, F. R. 2023. Honesty Is the Best Policy: Defining and Mitigating AI Deception. *NeurIPS 2023*.
- Ward, F. R. 2024a. Evaluating Language Model Character Traits.
- Ward, F. R. 2024b. The Reasons that Agents Act: Intention and Instrumental Goals.
- Wei, J. 2022. Emergent Abilities of Large Language Models.
- Weller, C. 2017. A robot has just been granted citizenship.
- Wooldridge, M. 1995. Intelligent agents: Theory and practice.
- Xi, Z. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey.
- Yin, Z. 2023. Do Large Language Models Know What They Don't Know?
- Yudkowsky, E. 2016. The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*.
- Yudkowsky, E. 2019. Coherent decisions imply consistent utilities.