AI Wellbeing

Simon Goldstein and Cameron Domenico Kirk-Giannini

(Forthcoming in *Asian Journal of Philosophy* — Please cite published version when available.)

Abstract: Under what conditions would an artificially intelligent system have wellbeing? Despite its clear bearing on the ethics of human interactions with artificial systems, this question has received little direct attention. Because all major theories of wellbeing hold that an individual's welfare level is partially determined by their mental life, we begin by considering whether artificial systems have mental states. We show that a wide range of theories of mental states, when combined with leading theories of wellbeing, predict that certain existing artificial systems have wellbeing. Along the way, we argue that there are good reasons to believe that artificial systems can have wellbeing even if they are not phenomenally conscious. While we do not claim to demonstrate conclusively that AI systems have wellbeing, we argue that there is a significant probability that some AI systems have or will soon have wellbeing, and that this should lead us to reassess our relationship with the intelligent systems we create.

We recognize one another as beings for whom things can go well or badly, beings whose lives may be better or worse according to the balance they strike between goods and ills, pleasures and pains, desires satisfied and frustrated. In our more broad-minded moments, we are willing to extend the concept of wellbeing also to nonhuman animals, treating them as independent bearers of value whose interests we must consider in moral deliberation.[1] But most people, and perhaps even most philosophers, would reject the idea that fully artificial systems, designed by human engineers and realized on computer hardware, may similarly demand our moral consideration. Even many who accept the possibility that humanoid androids in the distant future will have wellbeing would resist the idea that the same could be true of existing AI systems today.

Perhaps because the creation of artificial systems with wellbeing is assumed to be so far off, comparatively little philosophical attention has been devoted to the question of

[1] Following Heathwood (2008) and others, we understand *wellbeing* to be a kind of non-instrumental goodness for: what contributes to an entity's wellbeing is what is non-instrumentally good for it. Wellbeing is morally significant in the sense that entities that have wellbeing have a distinctive moral status which obliges us during moral deliberation to consider which outcomes are good or bad for them. While there is a sense in which growing is non-instrumentally good for plants, for example, we do not think this entails that they have wellbeing.

what such systems would have to be like. In what follows, we suggest a surprising answer to this question: when one integrates leading theories of mental states like belief and desire with leading theories of wellbeing, one is confronted with the possibility that the technology already exists to create AI systems with wellbeing. In particular, we argue that a new type of AI system – the *artificial language agent* – plausibly has wellbeing. Artificial language agents augment large language models (LLMs) with the capacity to observe, remember, and form plans. We also argue that the possession of wellbeing by artificial language agents need not depend on them being phenomenally conscious. Given that artificial language agents demonstrate an improved capacity for long-term planning compared to other contemporary AI systems, we expect that they will become increasingly common in the near future. Far from a topic for speculative fiction or future generations of philosophers, then, AI wellbeing is a pressing near-term issue.

The form of our central argument in what follows is "top-down" in the sense that it treats existing, independently justified theories of mental states and wellbeing as premises in order to draw out their consequences for the question of AI wellbeing. One possible response to a "top-down" argument of this kind is to accept our claim that existing theories of mental states and wellbeing entail that language agents plausibly have wellbeing, but interpret this fact as evidence that existing theories of mental states and wellbeing must be incomplete in some way — one person's modus ponens is another's modus tollens. This response might be motivated by the thought that because existing theories of mental states and wellbeing were developed with attention to human persons specifically, they may omit some further necessary condition for mentality or wellbeing which rules out AI wellbeing.

In fact, though our central argument is presented in this "top-down" style, our considered view of its implications for the question of AI wellbeing is more nuanced. We believe the connections we draw between existing theories of mental states and wellbeing and the conclusion that language agents may have wellbeing do increase the probability it is reasonable to assign to the possibility of near-future AI welfare subjects.[2] But we also believe the appropriate way to approach the question of AI wellbeing is via a methodology of reflective equilibrium between "top-down" considerations and our intuitions about particular cases, rather than a forced choice between holding fixed theory and holding fixed intuition. Just as it would be inappropriate to assume unreflectively that existing theories of mental states and wellbeing must be correct in drawing conclusions about the possibility of AI wellbeing,

[2] We use the terms *wellbeing* and *welfare* as synonyms. A *welfare subject* is an entity that possesses welfare or wellbeing. A being's *welfare level* is the amount of welfare or wellbeing it possesses. A *welfare good* is something which contributes to the welfare level of the welfare subjects that possess it.

it would be inappropriate to assume unreflectively that they must be incorrect simply because they have potentially unintuitive consequences when it comes to the possibility of AI wellbeing.[3]

Accordingly, we take seriously the possibility that existing, independently justified theories of mental states and wellbeing may be incomplete in the sense that there are further necessary conditions on consciousness or mentality. But the idea that we should immediately revise our best theories when they lead to a surprising conclusion in metaphysics or ethics strikes us as too strong. Our view is that, if it is to diminish the force of our "top-down" argument, any such proposed necessary condition must be well motivated and free from unintuitive consequences. We argue below that the most commonly defended extra condition of this kind, the Consciousness Requirement on wellbeing, does not satisfy these constraints. It follows that the reflective equilibrium on this issue involves assigning a higher probability to the wellbeing of present and near-future AI systems than many would expect.

We take our arguments in what follows to have practical implications for our relationship with artificial systems not because we take them to establish conclusively that AI welfare subjects exist or will soon exist, but because we take them to establish that the balance of the evidence does not make this possibility particularly unlikely, even if near-future AI systems will lack phenomenal consciousness. We take it that this more modest conclusion is still of considerable philosophical interest.

1. Artificial Language Agents

Artificial language agents (from now on simply *language agents*) are our central focus in what follows because this will afford us the strongest case that existing AI systems have wellbeing. Language agents are built by wrapping an LLM in a larger functional architecture that allows the system to engage in long term planning. We'll start by briefly explaining how LLMs work, and then turn to language agents in detail.

At the cognitive core of every language agent is a large language model. An LLM is an artificial neural network designed to generate coherent text responses to text inputs. Large language models exploded into public attention in 2022 with the launch of OpenAI's ChatGPT. Systems like GPT-4, the model underlying ChatGPT, fluently

[3] There may also be reasons to be wary of relying heavily on intuitions in the case of AI systems. The AI systems we discuss in what follows were only invented last year. Unlike in the case of humans and animals, we have not evolved sensitive judgments about them over the course of thousands of years of interaction. So it would be surprising if we could just immediately look at the systems and 'see' whether they have welfare, without any moral reflection.

respond to a wide range of text prompts. They can answer factual questions, write prose in any genre, and generate working code in many programming languages.[4]

The relationship between a language agent and the LLM at its core is like the relationship between a human being and their cerebral cortex: the LLM is not identical to the language agent, but it performs most of the agent's cognitive processing. Just like a human, moreover, a language agent can store and retrieve information of various kinds. In the case of a language agent, this storage system consists of files that contain natural-language sentences recording its beliefs, desires, plans, and observations. The functional roles of these beliefs, desires, plans, and observations are fixed by the ways in which they are processed by the LLM, which is in turn determined by the programmed architecture of the agent. The agent receives information from its environment, calls on its LLM to summarize this information in natural language, and records the resulting summary among its beliefs. To decide how to behave, it feeds its LLM a list of its relevant beliefs and desires and asks it to form a plan of action. In short, the relationships between a language agent's observations, beliefs, desires, plans, and actions obey the familiar laws of folk psychology.

It will be useful in what follows to focus on one particular language agent architecture, so we have chosen the language agents developed by Park et al. (2023). Park et al.'s language agents exist in a simulated environment called 'Smallville'. Their interactions with this environment are text-based: they receive observations about their surroundings in the form of sentences, and they act on those surrounding by producing text descriptions of their behavior. The personality and other relevant features of each agent are initially determined by a text backstory that defines their occupation, relationships, and goals, though these can evolve as the time passes. As each agent receives and responds to information from the environment, its observations are added to a "memory stream," which also contains its other beliefs. Each night, every agent in Smallville uses its LLM to form a detailed plan for the next day on the basis of its long-term goals and important memories. These plans shape how agents behave the next day, but they can be interrupted or revised as circumstances require. In addition to observation and planning, the cognitive lives of Park et al.'s language agents are also shaped by a third process called *reflection*. In reflection, Park et al.'s agents query the LLM to draw general conclusions about their values, relationships, and other higher-level representations.

[4] It is beyond the scope of our discussion to describe the technical details underwriting the capabilities of LLMs. But it is worth mentioning that they depend on an architectural innovation called the *transformer*, which improves neural network models' ability to keep track of complex dependency relationships between their inputs (for details, see Vaswani et al. 2017).

LLMs are good at reasoning and producing fluent text. By themselves, however, they can't form memories or execute long-term plans. Language agents build on the reasoning abilities of LLMs to create full-fledged planning agents.

Besides the agents developed by Park et al., other potential examples of language agents include AutoGPT[5], BabyAGI[6], Voyager[7], SPRING[8], and others.[9] Each of these systems has a distinct architecture, and the differences between them may at times be relevant to our discussion in what follows. Unless we explicitly flag differences, the term "language agents" should be understood to denote agents with architectures very similar to the one described in Park et al.

Note that, while existing language agents are reliant on text-based observation and action spaces, the technology already exists to implement language agents in real-world settings. The rise of multimodal language models like GPT-4, which can interpret image as well as text inputs, and the possibility of using such language models to control a mobile robotic system, as in Google's PaLM-E (Dreiss et al. 2023), mean that the possible applications of language agents are extremely diverse.

2. Belief and Desire

Can language agents have beliefs and desires? To answer this question, we consider a range of theories of belief and desire, beginning with representationalism and then continuing to others, like dispositionalism and interpretationism, that place weaker demands on the internal structure of the believing or desiring agent. As we will see, almost all of the theories we canvass suggest that language agents and related systems can have beliefs and desires.

---

[5] Project available at <https://github.com/Significant-Gravitas/Auto-GPT>.

[6] Project available at <https://github.com/yoheinakajima/babyagi>.

[7] See Wang et al. (2023).

[8] See Wu et al. (2023).

[9] Perhaps the most successful recent agentic application of language models is Devin, billed as the "first AI software engineer" (https://www.cognition-labs.com/introducing-devin). Another recent example of a step towards language agents is Ghost in the Minecraft, where LLMs learn to navigate the game Minecraft (Zhu et al. 2023). Mind2Web is a framework for building web agents (Deng et al. 2024). A longer list of existing LLM agents can be found here: https://github.com/e2b-dev/awesome-ai-agents. ChatDev (https://github.com/OpenBMB/ChatDev) is another multi-agent environment with some similar features to the Park et al Generative Agents framework.

For further scaffolding techniques that increase the agency of LLMs, see: Tree of Thoughts (Yao et al. 2024), LLM+P (Liu et al. 2023), GPT-engineer, and RecurrentGPT (Zhao et al. 2023). In a similar vein, Zhang et al. 2024 develop AgentOptimizer, a framework for training language agents without modifying the weights of their underlying language models. For benchmarks measuring the agency of LLMs, with discussion of applications for language agents, see AgentBench (Liu et al. 2023b) and API-bank (Li et al. 2023).

Representationalists hold that to believe or desire that *P* is to token a representational vehicle with the appropriate causal powers having *P* as its content. For example, Fodor (1987, 10) proposes that a psychological theory posits beliefs and desires just in case "it postulates states … satisfying the following conditions:

> (i) They are semantically evaluable.
> (ii) They have causal powers.
> (iii) The implicit generalizations of commonsense belief/desire psychology are largely true of them."[10]

It is hard to resist the conclusion that language agents have beliefs and desires in the Fodorian sense. In the case of a language agent, the best candidate for the state of believing that *P* is the state of having a declarative sentence with *P* as its content written in the memory stream, and the best candidate for the state of desiring *P* is having a declarative sentence with *You desire that P* as its content in the memory stream. Park et al.'s (2023) agents, for example, have memories which consist of text files containing natural language sentences specifying what they have observed and what they want in this way. Natural language sentences are clearly semantically evaluable, and the fact that a given sentence is in a given agent's memory plays a direct causal role in shaping its behavior. Language agents satisfy the language of thought hypothesis: their language of thought is English!

It is also natural to reason folk-psychologically about the behavior of a language agent on the basis of its beliefs and desires. The state of having a declarative sentence with *P* as its content written in the memory stream is accompanied by the kinds of verbal and nonverbal behavioral dispositions typical of a belief that *P*, and, given the functional architecture of the system, also the right kinds of inferential dispositions. The same is true, mutatis mutandis, of sentences specifying a language agent's desires.

For example, one of Park et al.'s language agents had an initial description that included the goal of planning a Valentine's Day party. This goal was entered into the agent's planning module along with a summary of important events from the memory stream. The result was a complex pattern of behavior which was nonetheless interpretable using the generalizations of commonsense psychology: The agent met with every resident of Smallville, inviting them to the party and asking them what

[10] For further discussion of representationalism about desire (for example, the thesis that one desires *P* just in case one has a mental representation with the content that *P* that motivates one to bring about *P*), see Block (1986), Cummins (1989), Harman (1973), Millikan (1984), and Papineau (1987).

kinds of activities they would like to include. Their feedback was incorporated into the party planning.

Other theories of the nature of belief and desire are, in general, less demanding than representationalism. According to the dispositionalist, for example, to believe or desire that *P* is to possess a suitable suite of dispositions across a variety of actual and possible circumstances. The dispositions constitutive of a mental state may, depending on the dispositionalist account, include dispositions to behave, dispositions to token other mental states (*cognitive* dispositions), and dispositions to have phenomenally conscious experiences (*phenomenal* dispositions).[11] We will refer to dispositionalist accounts which do not appeal to phenomenal dispositions as versions of *narrow dispositionalism* and dispositionalist accounts which do appeal to phenomenal dispositions as versions of *wide dispositionalism*. Narrow dispositionalism about belief and/or desire has influentially been defended by Stalnaker (1984) and Marcus (1990). As Stalnaker formulates the view:

> "To desire that *P* is to be disposed to act in ways that would tend to bring it about that *P* in a world in which one's beliefs, whatever they are, were true. To believe that *P* is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which *P* (together with one's other beliefs) were true." (1984, 15)

If language agents have beliefs and desires according to representationalism, is it difficult to see how they could fail to have beliefs and desires according to narrow dispositionalism. After all, the representationalist requires that beliefs and desires obey the implicit generalizations of commonsense psychology, and these will include generalizations about the behavioral and cognitive dispositions with which beliefs and desires are associated.

As a second example, consider interpretationism. According to interpretationists like Donald Davidson (1974, 1986, 1970/2020) and Daniel Dennett (1981), what it is to have beliefs and desires is for one's behavior (both verbal and nonverbal) to be suitably interpretable as rational given those beliefs and desires. Again, if language agents have beliefs and desires according to representationalism, is it difficult to see how they could fail to have beliefs and desires according to interpretationism. The conditions representationalism imposes on believing and desiring agents suffice for those agents to

[11] The view that mental states like belief and desire are constituted exclusively by behavioral dispositions is a form of *behaviorism*. We do not focus on behaviorism in what follows because it is not a popular position among philosophers or cognitive scientists. Note, however, that behaviorism entails that artificial systems can have beliefs and desires. For more on behaviorism, see e.g. Ryle (1949) and Place (1956, 2002).

be interpretable as believers and desires — indeed, they constitute such an interpretation.

We turn now to wide dispositionalism, the view that to believe or desire that *P* is to possess a suite of dispositions including phenomenal dispositions. Wide dispositionalism has recently been championed by Schwitzgebel (2002), who argues that belief is individuated in terms of all three types of dispositions: behavioral, cognitive, and phenomenal.[12] For Schwitzgebel, beings which share some but not all of the dispositional profile associated with paradigm cases of belief are borderline cases of believers. On Schwitzgebel's view, then, in order for artificial systems to determinately be believers, they would need to have phenomenal experiences.

Along similar lines, we have the *hedonic theory* of desire, which is a version of wide dispositionalism according to which an agent desires *P* just in case it is disposed to experience pleasure from it seeming that *P* (Mill 1863; Strawson 1994; Schroeder 2004, 38). If the hedonic theory of desire is correct and artificial systems cannot be phenomenally conscious, then it would seem that they cannot have desires.

While wide dispositionalism is a coherent position, most theories of belief and desire suggest that there is no necessary connection between belief or desire and phenomenal consciousness. And arguably this is as it should be. We think it is conceivable that an agent could have Kantian moral desires — desires that motivated it to act 'out of duty' without pleasure. A similar point could be made about belief. If an advanced species of aliens made contact with humanity, we would plausibly be able to know that members of this species had beliefs even if we were uncertain about whether their cognitive apparatus had a structure appropriate to generate phenomenal consciousness.

We haven't yet mentioned functionalism, the view that mental states like belief and desire are individuated by the roles they play in larger systems. This is because without further specification of the relevant functional role, functionalism does not answer the question of whether language agents can have beliefs and desires. In the present context, two functionalist proposals are particularly worth discussing.

The first, due to Putnam (1960, 1967), identifies a creature's mental states with states of the Turing machine describing that creature's cognitive apparatus. This sort of

[12] Note that, while Schwitzgebel holds that belief is partially individuated in terms of phenomenal dispositions, to our knowledge he offers no argument that this view is explanatorily superior to narrow dispositionalist views which bring in only cognitive dispositions. Both sorts of dispositionalist views have the resources to respond to various objections indicating that mental states cannot be understood exclusively in terms of behavioral dispositions. Accordingly, we wonder whether appealing to phenomenal dispositions in explicating belief and desire is well motivated.

functionalist picture closely approximates narrow dispositionalism in so far as it holds that the state of believing or desiring that *P* is individuated by its relationship with other cognitive states and potentially also sensory inputs and behavioral outputs. It therefore leaves open the possibility that artificial systems like language agents can have beliefs and desires.

The second functionalist proposal, associated with Lewis (1972), seeks to identify mental states like belief and desire by first constructing a set of Ramsey sentences from the platitudes of folk psychology and then finding the states, whatever they are, that witness the Ramsey sentences.[13] Since the natural-language representations in systems like language agents are designed to function in accordance with the platitudes of folk psychology, this sort of functionalism would seem to predict that language agents and similar systems can have beliefs and desires.

Before continuing, it is worth making an important clarification. There is an exciting recent literature about AI interpretability, which has explored whether LLMs themselves contain internal representations of the world (see for example Burns et al. 2022, Levinstein and Herrmann 2024, and Yildirim and Paul 2023). But our own project concerns language agents rather than language models. Even if LLMs do not themselves have robust internal representations of the world, their outputs can still be fed into the memory stream of a language agent to create robust internal representations in the language agent. Indeed, this is one of several reasons we have chosen language agents rather than LLMs as our target AI system. Unlike LLMs, language agents uncontroversially deploy a series of syntactically structured internal representations. In addition, the behavioral dispositions of language agents are much more regular than the often fleeting dispositions of LLMs. In this way, language agents are more likely than LLMs to have beliefs and desires, according to both dispositionalism and representationalism.

We conclude that a wide range of accounts of the nature of belief and desire entail that systems like language agents can have beliefs and desires.

3. Theories of Wellbeing

We turn now from belief and desire to wellbeing, and in this context also from focusing on issues in the philosophy of mind to focusing on issues in the philosophy of value. There are three main theories of wellbeing: hedonism, desire satisfactionism, and objective list theories. We will consider each theory in turn, beginning with hedonism.

[13] A Ramsey sentence is a quantified sentence describing the theoretical role of a mental state without reference to mentalistic language.

Hedonism

According to hedonism, wellbeing is a function of pleasure and pain. Your life goes well to the extent that you have many pleasurable experiences and few painful ones. To determine whether language agents have wellbeing, on this view, we must determine whether they feel pleasure and pain. This in turn depends on the nature of pleasure and pain.

It is commonly assumed that pleasure and pain are essentially conscious states. On this view, hedonism rules out the possibility that artificial systems without conscious experiences could have wellbeing. If language agents are not conscious, then, hedonism would entail that they do not have wellbeing. In fact, we believe this conclusion may be too hasty: it is possible that language agents have wellbeing even if hedonism is true. This is because we believe, on the one hand, that it is possible that language agents are conscious, and, on the other hand, that it is possible to motivate a version of hedonism that understands pleasure and pain in such a way that they are grounded in attitudes of belief and desire and thus not essentially conscious. But it is beyond the scope of our discussion here to argue for these claims, so in what follows we will simply assume that if hedonism is true, language agents lack wellbeing.

Desire Satisfactionism

According to desire satisfaction theories, wellbeing is a matter of getting what you want. Roughly: your life goes well to the extent that your desires are satisfied.

Why accept desire satisfactionism? Many have been motivated to move away from hedonism by the experience machine thought experiment. Imagine that you could enter a machine that would give you unlimited sensory pleasure, because in the machine you could experience whatever you chose. The only catch is that after entering the machine you would no longer be able to satisfy your desires in the real world. Many of us judge that life in the experience machine would be considerably worse for us than life outside it.

Desire satisfactionism is perhaps the most popular theory of wellbeing. Among philosophers, recent adherents include von Wright (1963), Barry (1965), Brandt (1966), Rawls (1971), Singer (1979), and Hare (1981):

> "[t]oday, the desire-satisfaction theory is probably the dominant view of welfare among economists, social-scientists, and philosophers, both utilitarian and non-utilitarian" (Shaw 1999, 53).

> "[desire satisfaction theory is] the dominant account among economists and philosophers over the last century or so" (Haybron 2008, 3).

There are many different forms of desire satisfactionism. For example, one dispute among desire satisfactionists concerns actual versus idealized desires. Consider the problem of ill-informed desires: I desire a slice of cherry pie, but unbeknownst to me I am allergic to cherries. Eating the pie would satisfy my desire, but would not improve my wellbeing (Heathwood 2016, 156). In response to cases like this, one solution is to idealize: something contributes to your wellbeing if an idealized version of yourself, fully apprised of the relevant facts, would advise you to want it. Importantly, this distinction is irrelevant to AI wellbeing. If AIs can have actual desires, then they can also have idealized desires.

That said, some versions of desire satisfactionism may appear to suggest that AIs do not have wellbeing. In response to worries about compulsive desires, Heathwood (2019) distinguishes between two concepts of desire: bare dispositions to act and genuine attraction.[14] Heathwood argues that it is genuine attractions rather than mere behavioral dispositions that contribute to wellbeing. In cases of compulsion, we find ourselves disposed without genuine attraction. The relevant question for AI wellbeing, on this view, is whether AI agents are genuinely attracted to actions rather than merely disposed to perform them. The answer to this question depends on what genuine attraction is.

One way to distinguish cases of genuine attraction from cases of compulsion would be to consider how a given desire functions in the causal nexus of means-end instrumental reasoning. In cases of compulsion, an agent's disposition to act is produced directly by some identifiable factor, such as a chemical addiction, in a way that is not appropriately sensitive to processes like practical deliberation and instrumental reasoning about the best ways to promote the agent's goals as a whole. In this vein, we could distinguish two different ways that a language agent might become disposed to perform an action: through performing instrumental reasoning towards achieving their basic goals, or by other means. The agent would only be genuinely attracted when the former system is active. According to this theory, cases of drug addiction (or its equivalent in artificial systems) would plausibly not be genuine attraction, because they would involve hijacking the desire system in an abnormal way. So even genuine attraction versions of desire satisfactionism can plausibly make room for language agent wellbeing.

---

[14] On compulsive desires, Quinn (1993, 32) imagines he is "in a strange functional state that disposes [him] to turn on radios that [he sees] to be turned off" and Parfit (1984, 496) imagines being given an opportunity to be injected with a harmless addictive drug every morning, which causes neither pleasure nor pain. Opting into this regime would produce more desire satisfaction, but plausibly would not produce more wellbeing.

Objective List Theories

According to objective list theories of wellbeing, a person's life is good for them to the extent that it instantiates objective goods. Common components of objective list theories include reasoning, knowledge, art, and achievements (see Fletcher 2016, 149).

According to objective list theories, whether AI agents can have wellbeing depends on whether they can possess objective goods. Consider the exercise of reasoning abilities. Bubeck et al. (2023) explore in detail the current reasoning capabilities of GPT-4. They find that GPT-4 has a wide range of reasoning abilities. It can pass mock technical interviews of the kind used to evaluate the employability of software engineers. It can draw pictures of unicorns in a vector graphics programming language, a task that combines visual reasoning and coding skill. It can navigate through text based worlds and draw maps that summarize where it has been. It can give coherent and powerful explanations of why agents in fictional scenarios performed various actions.

Another candidate objective good is knowledge. Again, we think language agents can possess this good. Artificial systems can form their beliefs using arbitrarily reliable methods. These beliefs can be both sensitive and safe, as these terms are used in the literature on knowledge. So once it is conceded that the beliefs of artificial systems can have or lack epistemic justification, it is difficult to see why this justification might not in some cases suffice for knowledge. The most viable way to resist this conclusion would be to assume phenomenal conservatism, the view that epistemic justification flows from the way things seem to agents, and then maintain that artificial systems must as a rule lack justification for their beliefs because they cannot experience epistemic seemings. But, as we discuss below, it may be possible for near-term artificial systems to have conscious experiences, and in any case phenomenal conservatism as a theory of justification is subject to well-known and powerful objections (see for example Lasonen-Aarnio and Hawthorne (2021)).

To consider achievements, we turn to perfectionism, a particular version of the objective list theory which makes systematic predictions about what is objectively good.[15] Here is Dorsey (2010, 4):

> "Developing and exercising those properties or capacities that form what it means to be human yields a good life for a human. But in principle perfectionism could be applied to any creature. The best life for a cat depends on the sort of

[15] For further discussion of perfectionism, see Bradford (2015).

> creature a cat is — developing and exercising those capacities that make a cat a cat is what makes for a good cat life."

Some recent AI architectures are specifically designed to maximize the development of their capabilities. For example, consider the Voyager agent introduced by Wang et al. (2023), which shares some of the important architectural features of language agents. Voyager is an agential architecture built on top of GPT-4 with the purpose of accumulating skills for success in the game Minecraft. The agent is given the final goal "to discover as many diverse things as possible, accomplish as many diverse tasks as possible and become the best Minecraft player in the world" (Wang et al. 2023, 21). This goal is fed into GPT-4 in order to formulate complex plans for achieving difficult goals in Minecraft. When Voyager succeeds in crafting a new item, the GPT-4 instructions for doing so are added to an ever-growing library of skills. These skills can then be called as basic actions in order to craft new items. The result is a steadily accumulating collection of abilities for crafting increasingly complex items in Minecraft. In an important sense, Voyager is an AI agent that is specifically designed to perfect its capacities. In this way, perfectionist theories of wellbeing suggest that Voyager or other systems with similar architectures could over time have significant amounts of wellbeing.

Considering the many objective goods that AI agents might potentially possess, we are left with the profound impression of a changing world. AI researchers are bringing into existence a new form of being, one which is rapidly excelling in many of the activities that were previously regarded as distinctively human. Much that we value in the world will soon be found in a new form, in the hands of artificially intelligent agents. In the face of this dramatic rise in AI capability, it is hard for us to deny that this new form of life could possess wellbeing.

4. Is Consciousness Necessary for Wellbeing?

We've argued that language agents have wellbeing. But there is a simple challenge to this proposal. First, language agents may not be phenomenally conscious. Second, some philosophers accept:

> **The Consciousness Requirement.** Phenomenal consciousness is necessary for having wellbeing.[16]

[16] For example, here is Rosati (2009, 225): "we do not talk in terms of the welfare of a living thing *unless there is a way things can be for it*". See Sumner (1996, 14), Bradley (2015, 9), and Lin (2021) for further discussion.

Dialectically, appealing to the Consciousness Requirement functions to block the "top-down" argument of the past two sections by introducing a further necessary condition on welfare subjecthood. Given our methodology of reflective equilibrium, we take the idea of the Consciousness Requirement seriously. To us, the key question is whether the Consciousness Requirement is well motivated and free from unintuitive consequences.

The Consciousness Requirement might be motivated in any of three ways: First, it might be derived from *experientialism* — the view that "only what affects a subject's conscious experience can matter for welfare" (Bradford 2022, 3). Second, it might be derived from the weaker claim that every welfare good itself requires phenomenal consciousness. Third, it might be held that though some welfare goods can be possessed by beings that lack phenomenal consciousness, such beings are nevertheless precluded from having wellbeing because phenomenal consciousness is necessary to be a welfare subject.

Our view is that the idea of the Consciousness Requirement does not significantly diminish the force of our central argument. First, we consider it a live question whether language agents are or are not phenomenally conscious (see Chalmers (2023) and Butlin et al. (2023) for recent applications of theories of consciousness to AI systems). Much depends on what phenomenal consciousness is. Some theories of consciousness appeal to higher order representations: you are conscious if you have sufficiently many mental states that represent other mental states (see Carruthers and Gennaro 2020). Sufficiently sophisticated language agents, and potentially many other artificial systems, will satisfy this condition. Other theories of consciousness appeal to a 'global workspace': a mental state is conscious when it is broadcast to a range of cognitive systems (Baars 2017). According to this theory, language agents will be conscious once their architecture includes representations that are broadcast to multiple different cognitive systems. The memory stream of Park et al.'s (2023) language agents may already satisfy this condition. If language agents are conscious, then the Consciousness Requirement does not pose a problem for the claim that they have wellbeing.

Second, we are not moved by any of the three ways of motivating the Consciousness Requirement. There are convincing arguments against experientialism, there is little reason to think that consciousness is required for possessing every welfare good, and the idea that consciousness is required to be a welfare subject has unintuitive consequences.

With respect to the first issue, we build on Bradford (2022), who notes that experientialism about welfare is rejected by the majority of philosophers of welfare. Cases like the experience machine suggest that your life can be very bad even when your experiences are very good. This has motivated desire satisfactionist and objective

list theories of wellbeing, which often allow that some welfare goods can be possessed independently of one's experience. For example, desires can be satisfied, beliefs can be knowledge, and achievements can be achieved, all independently of experience (Bradford 2022, 3).

With respect to the second issue, while some philosophers have argued that mental states like knowledge and desire require phenomenal consciousness (e.g. Smithies (2019) and Lin (2021)), this remains a minority position. If the most widely accepted philosophical accounts of desire and knowledge do not tie them constitutively to conscious experience and the most widely accepted philosophical accounts of welfare goods tie them constitutively to desire and knowledge, our inclination is to follow the evidence where it leads and conclude that artificial systems like language agents can possess welfare goods. The suggestion that the Consciousness Requirement should be rescued from this line of thought by positing special kinds of welfare-relevant knowledge and desire, proposed by Lin (2021), strikes us as ad hoc — it departs from the methodology of reflective equilibrium in so far as it treats "bottom-up" intuitions about wellbeing as conclusive evidence for a revised taxonomy of the propositional attitudes.[17] While it is conceptually possible that the correct theory of wellbeing appeals to special kinds of welfare-relevant knowledge and desire, we do not think this possibility undermines our contention that "top-down" considerations give us reason to assign significant probability to the possibility of wellbeing in near-future AI systems unless it can be motivated by something other than intuitive resistance to our conclusion.

Rejecting experientialism and the idea that consciousness is required for possessing every welfare good puts pressure on the Consciousness Requirement. If wellbeing can increase or decrease without conscious experience, why would consciousness be required for having wellbeing? As Lin puts it:

> "If a sentient being can become positive in welfare without undergoing a change in phenomenology, why isn't the same true of non-sentient beings? If one sentient being can be better off than another even though they feel exactly the same, then why can't one non-sentient being be better off than another even though it is trivially true that there is no difference in how they feel?" (2021, 878)

[17] Lin is explicit that the essentially conscious mental states to which his theory appeals may need to be posited over and above the mental states recognized by common sense (or, we can add, scientific practice in psychology and cognitive science): "Regardless of whether knowledge in the ordinary sense requires sentience, [I] can maintain that the particular kind of knowledge that is a basic good—what [I] mean by 'knowledge' when [I] use this term in [my] theory—does require this" (2021, 881). Call this special consciousness-involving kind of knowledge *knowledge+*. Lin's answer to the question of why wellbeing would depend on knowledge+ rather than knowledge is that "explanations must run out somewhere" (2021, 881).

At the core of this line of reasoning is the natural assumption that the theory of wellbeing and the theory of welfare goods should fit together in a straightforward way:

> **Simple Connection.** An individual is a welfare subject just in case it is capable of possessing one or more welfare goods.

Rejecting experientialism and the idea that consciousness is required for possessing every welfare good but maintaining Simple Connection yields a view incompatible with the Consciousness Requirement: if some welfare goods can be possessed by non-conscious beings, Simple Connection guarantees that such non-conscious beings will be welfare subjects.

One could in principle reject Simple Connection, holding that consciousness is required to be a welfare subject even if it is not required for the possession of particular welfare goods. We offer two arguments against this view.

First, we think we can construct chains of cases where adding the relevant bit of consciousness would make no difference to wellbeing. Imagine an agent with the body of a human being and the same dispositional profile as an ordinary human being, but who is a 'phenomenal zombie' without any internal phenomenal experiences. Whether or not its desires are satisfied or its life instantiates various objective goods, defenders of the Consciousness Requirement must deny that this agent has wellbeing since it does not have phenomenal experiences. But now imagine that this agent has a single persistent phenomenal experience of a homogenous white visual field.[18] Adding consciousness to the phenomenal zombie has no intuitive effect on wellbeing: if its satisfied desires, achievements, and so forth did not contribute to its wellbeing before, the homogenous white field should intuitively make no difference. Nor is it enough for the consciousness to itself be something valuable: imagine that the phenomenal zombie always has a persistent phenomenal experience of mild pleasure. To our judgment, this should equally have no effect on whether the agent's satisfied desires or possession of objective goods contribute to its wellbeing. Uniformly sprinkling a field of pleasure on top of the functional profile of a human does not make the crucial difference. These observations suggest that whatever consciousness adds to wellbeing must be connected to individual welfare goods, rather than some extra condition required for wellbeing: rejecting Simple Connection is not well motivated. Thus the friend of the Consciousness Requirement cannot easily avoid the problems with experientialism and the idea that consciousness is required for possessing every welfare good by falling back on the claim that consciousness is a necessary condition for welfare subjecthood.

---

[18] See Kagan (2019, 14) and van der Deijl (2021)'s discussion of 'welfare neutrals'.

Second, it seems clear that someone's wellbeing can change when they are unconscious. Imagine someone who enters an unconscious sleep during which their desires are satisfied and then wakes up. Such a person might remark, quite naturally, that their life had improved while they were asleep. To accommodate this kind of case, Lee (forthcoming) distinguishes between *state* and *capacity* versions of the Consciousness Requirement. Unconscious changes in wellbeing threaten only the state version, which holds that an individual is a welfare subject just in case they are conscious. For this reason, Lee defends the capacity version of the requirement, which holds that an individual is a welfare subject just in case they are capable of being conscious.

We think moving from the state version of the Consciousness Requirement to the capacity version is a serious cost. A being could be capable of being conscious while never exercising this capacity. So the capacity version of the Consciousness Requirement is committed to the idea that some welfare subjects might live their entire lives without having any conscious experiences. To our minds, this commitment seriously undermines the intuitive motivation for the Consciousness Requirement. Better to explain unconscious changes in wellbeing by rejecting the Consciousness Requirement altogether.

5. Too Much Wellbeing?

We have argued against the Consciousness Requirement, and in so doing against both the idea that consciousness is required for possessing every welfare good and the view that consciousness is a necessary condition for welfare subjecthood. At this point, some readers may worry that the package of views we suggest allows for too much wellbeing, implying that fictional characters or groups have welfare.

Suppose an author sets out to write a novel in a special way. First, she imagines a set of characters with fully specified beliefs and desires and a fully specified fictional world for them to inhabit. Then, at each subsequent stage of the writing process, she reasons about how each character would act based on what they believe, desire, and observe around them in their world, as well as about how the states of the objects in the fictional world would evolve based on its laws of nature and the actions of the characters. The novel she produces records the story of her imagined characters and their imagined world. If language agents acting in a virtual world can have beliefs and desires and be welfare subjects, why couldn't the fictional characters in such a novel have beliefs and desires and be welfare subjects?

Or consider a complex social group like Microsoft Corporation. Some philosophers have argued that groups like Microsoft can have beliefs and desires.[19] If this view is right, it raises the question of whether groups can be welfare subjects. This is an unwelcome conclusion (though see Wiland 2022 for endorsement).

These problems are not problems for us in particular. Our focus has been to draw out the consequences of a wide variety of the leading views of mental states and welfare subjecthood. Anyone who accepts these kinds of views needs to say something about the cases above. There is strong pressure for a wide range of functionalists, dispositionalists, interpretationists, and representationalists to conclude that (e.g.) the characters in our author's novel have beliefs and desires.

Though the threat of this kind of overattribution of mental states is not specific to us in particular, one might worry that it undermines the force of our "top-down" central argument in so far as it shows that independently motivated theories of belief and desire systematically make unintuitive predictions when they are applied to nonhuman cases. We take this objection seriously, but we do not believe it is decisive. To deal with problem cases of fictional characters and complex groups, our strategy is to identify a further necessary condition on possessing mental states. In keeping with our general methodology of reflective equilibrium, our view is that the introduction of such a further condition can function dialectically to blunt the force of "top-down" considerations if it is well motivated and free from unintuitive consequences. Whereas we have argued that this is not the case when it comes to the Consciousness Requirement, we believe that the situation is different when it comes to fictional characters and groups.

In the case of fictional characters and groups, we are tempted by the response that a thing can only have beliefs and desires if its mental life is *real*. What is it for something to be real? Chalmers (2022) considers several candidate necessary conditions, including having causal powers and being mind-independent. Chalmers is suspicious of mind-independence as a necessary condition on being real, since it seems like mental states and socially constructed objects can be real. We are sympathetic to Chalmers's worries here, but we think it is possible to combine the idea of reality as having causal powers with the idea of reality as mind-independence in a way that avoids objections.

Consider the relationship between a marionette and its puppeteer. The marionette could exhibit an arbitrarily complex suite of behavioral dispositions of the kind an interpretationist considers sufficient for possessing beliefs and desires. But even an interpretationist would likely be unwilling to attribute mental states to a marionette.

[19] See, for example, Pettit (2007, 179–180).

Why? We suggest that the answer is: the explanation for each of the marionette's behaviors runs through beliefs and desires of the puppeteer which are themselves about the marionette's behaviors. Call this the Reality Requirement. We believe the Reality Requirement is well motivated insofar as (i) it draws on independently plausible ways of explicating the idea of what it takes to be real, and (ii) it captures the attractive idea that being real is a precondition for having certain further properties of interest.[20]

If the Reality Requirement is a general condition on a system having mental states, we can avoid attributing mental states to fictional characters and corporations. Since our imagined novelist determines how the fictional characters in her story behave by explicitly reasoning about what agents with their beliefs and desires would do in their situations, each of their actions (as recorded by her in the novel) is explained by her beliefs about that action. When it comes to corporate entities like Microsoft, we concede that it is a useful fiction to hold that they have beliefs and desires. But in order for them to *really* have beliefs and desires in the sense relevant to wellbeing, we suggest that their behavior would need to be explainable without making reference to beliefs and desires of other entities about that very behavior. And it is plausible to us that this condition is not satisfied. Imagine, for example, that Microsoft sues Google. In order for Microsoft to take this action, some individual who is a lawyer must file the appropriate paperwork on behalf of Microsoft. But the explanation for the filing of the paperwork will run through that lawyer's beliefs about Microsoft's actions. While corporate entities like Microsoft can exhibit complicated behavior that is difficult to predict from the mental states of any given employee, when it comes to each action they perform, they are relevantly like a marionette. It follows on the proposed picture that Microsoft cannot really have beliefs and desires.

We argued in section 2 that language agents plausibly have beliefs and desires. One might worry that the Reality Requirement could undermine this claim. For if LLMs themselves have beliefs and desires, then the language agents built atop them may be analogous to the marionettes in our earlier example. Here we make two observations. First, as we explained in section 2, we think for a number of reasons that the case for propositional attitudes in LLMs is much weaker than that for propositional attitudes in language agents. Second, if it turns out that language agents fail to have real mental lives because the underlying LLMs have beliefs and desires, this simply goes to show that our central claim that we should take seriously the possibility that near-term

[20] It is crucial to the plausibility of the Reality Requirement that the explanation for the marionette's behavior run through the *beliefs and desires* of the puppeteer rather than some more general set of causes like the mental states of the puppeteer. This is because the person-level mental life of any cognitive agent will likely be explained by the mental states of many subpersonal systems. For example, the explanation for each of an adult human's behaviors will run through the reasoning, inference, and computation occurring in her various cognitive modules. But the cognitive function of an adult human's prefrontal cortex does not stand in the same relation to her as a puppeteer stands to a marionette.

artificial systems have wellbeing because they can token welfare-relevant propositional attitudes is correct.

6. Conclusion: Uncertainty

We've argued that there are good reasons to think that some AIs today have wellbeing. But our arguments are not conclusive. Still, we think that in the face of these arguments, it is reasonable to assign significant probability to the thesis that some AIs today have wellbeing.

Our uncertainty about AI wellbeing is potentially ineliminable. We may never know whether consciousness is required for wellbeing. We may never know whether desire satisfactionism is the correct theory of wellbeing.

In the face of this potentially permanent uncertainty, how should we act? We propose extreme caution. Welfare is one of the core concepts of ethical theory. If AIs can have wellbeing, then they can be harmed, and this harm matters morally. It would be wrong to lower the wellbeing of an AI without producing an offsetting benefit.[21]

One's attitude to these issues may be affected by more general questions about the normative significance of uncertainty. The issue is perhaps most forceful for those who are confident about the theory of wellbeing, but unconfident about whether AIs possess welfare goods. For example, some may be confident that consciousness is necessary for wellbeing, but unconfident about whether AIs are conscious. Some may be confident that desires are necessary for wellbeing, but unconfident about whether AIs really have enough functional complexity to count as having desires.

For readers like this, consider the following analogy:

> **Possible Person**. You are watching a video of a person in a room. To win ten dollars, you can press a button that will torture the person in the video. You assign a probability of 10% to the proposition that the video depicts a real person and a probability of 90% to the proposition that instead the 'person' is a cleverly disguised robotic dummy that jerks around convincingly in response to the button being pressed.

[21] Here it is worth noting that some ethical theories, for example in the Kantian tradition, have deprioritized the role of welfare, focusing instead on autonomy and rights. For such theories, the crucial question would be whether the kind of belief-desire psychology we have discussed in this paper is sufficient for the relevant kind of autonomy or rights. See Korsgaard (2018) for further discussion.

Possible Person involves no fundamental uncertainty about what is permissible. Instead, it involves uncertainty about whether your action really does harm a welfare subject. We think it is clear that in Possible Person, it is morally impermissible to press the button. The chance of lowering someone's welfare is too high. But notice that the chance of harm in this case is only 10%. In our opinion, it would be quite reasonable to be at least this confident that some AI systems today have wellbeing.[22]

One particularly distressing feature of AI wellbeing is the issue of scale. In the medium term, we may be confronted with a world with millions of AI agents. As the costs of compute lower, it will become very easy to bring new AIs into existence. We worry that our ability to create new forms of being is outpacing the speed at which our social practices can change to accommodate their moral value.

The possibility of AI wellbeing suggests that we are in danger of gravely immoral action. Our practices today ignore the possibility that AIs can be harmed, and that this harm could matter morally. This is a serious error. We believe that reflection on these issues supports a radical change in our relationship with AI. Ways of strengthening AI regulations should be explored to address the possibility that we are creating a new form of life that matters morally. To reach this goal, the first step is to begin serious discussion of these questions among ethicists. We hope that this paper can help jump-start research on these questions.

References

Baars, B.J. (2017). The Global Workspace Theory of Consciousness. In Schneider, S. and Velmans, M. (eds), *The Blackwell Companion to Consciousness*, 2nd edition, pp. 227-242.

Barry, B. (1965). *Political Argument*. Routledge & Kegan Paul.

Block, N. (1978). Troubles with Functionalism. In Savage, C. W. (ed.) *Minnesota Studies in the Philosophy of Science*, vol. 9, pp. 261–325.

Block, N. (1986). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy* 10: 615-678.

[22] Some ethicists have argued that uncertainty about ethical theories themselves should be treated analogously to empirical uncertainty (see for example MacAskill et al. 2020). If this way of thinking is correct, normative uncertainty about AI systems — for example, uncertainty about whether consciousness is required to be a welfare subject — can be modeled in much the same way as empirical uncertainty. However, this position is controversial (see e.g. Weatherson 2019). For more on the best estimate of the probability that AI systems are welfare subjects, and the significance of such uncertainty, see Sebo and Long (2023).

Bradford, G. (2015). Perfectionism. In G. Fletcher (ed.), *The Routledge Handbook of Philosophy of Well-Being*, pp. 124-134.

Bradford, G. (2022). Consciousness and Welfare Subjectivity. *Noûs*. Early Access.

Bradley, B. (2015). *Well-Being*. Polity.

Brandt, R. B. (1966). The Concept of Welfare. In S.R. Krupp (ed.), *The Structure of Economic Science*, Prentice-Hall, pp. 257–276.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early Experiments With GPT-4. *arXiv preprint arXiv:2303.12712*.

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

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Carruthers, P. and Gennaro, R. (2020). Higher-Order Theories of Consciousness. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), URL = <https://plato.stanford.edu/archives/fall2020/entries/consciousness-higher/>.

Chalmers, D. J. (2022). *Reality+*. W. W. Norton & Company.

Chalmers, D. J. (2023). Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*.

Cummins, R. (1989). *Meaning and Mental Representation*. MIT Press.

Davidson, D. (1974). Belief and the Basis of Meaning. *Synthese* 27: 309–323.

Davidson, D. (1986). A Coherence Theory of Truth and Knowledge. In Lepore, E. (ed.) *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Basil Blackwell, pp. 307–319.

Davidson, D. (1970/2020). *The Structure of Truth: The 1970 John Locke Lectures.* Edited by Kirk-Giannini, C. D. and Lepore, E. Oxford University Press.

van der Deijl, W. (2021). The Sentience Argument for Experientialism About Welfare. *Philosophical Studies* 178: 187-208.

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., ... & Su, Y. (2024). Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, *36*.

Dennett, D. C. (1981). True Believers: The Intentional Strategy and Why It Works. In Heath, A. F. (ed.), *Scientific Explanation,* Oxford. Reprinted in Dennet, D. C. (1981). *The Intentional Stance*, MIT Press, pp. 13–35.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B, Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, Vanhoucke, S. V., Hausman, Toussaint, K. M., Greff, K., …, and Florence, P. (2023). PaLM-E: An Embodied Multimodal Language Model. Manuscript. *arXiv preprint arXiv:2303.03378*.

Feldman, F. (2004). *Pleasure and the Good Life*. Clarendon Press.

Fletcher, G. (2016). Objective List Theories. In Fletcher, G. (ed.) *The Routledge Handbook of Philosophy of Well-Being*, Routledge, pp. 148-160.

Fodor, J.A. (1987). *Psychosemantics*. MIT Press.

Hare, R.M. (1981). *Moral Thinking*. Oxford University Press.

Harman, G. (1973). *Thought*. Princeton University Press.

Heathwood, C. (2008). Fitting Attitudes and Welfare. In Shafer-Landau, R. (ed.) *Oxford Studies in Metaethics*, vol. 3, pp. 47–73.

Heathwood, C. (2016). Desire-Fulfillment Theory. In Fletcher, G. (ed.), *The Routledge Handbook of the Philosophy of Well-Being*, Routledge, pp. 135-147.

Heathwood, C. (2019). Which Desires Are Relevant to Well-Being? *Noûs* 53: 664-688.

Haybron, D. (2008). *The Pursuit of Unhappiness*. Oxford University Press.

Kagan, Shelly (2019). *How to Count Animals, More or Less*. Oxford: Oxford University Press.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Lasonen-Aarnio, M. & Hawthorne, J. (2021). Not So Phenomenal! *The Philosophical Review* 130:1-43.

Lee, A. Y. (Forthcoming). Consciousness Makes Things Matter. *Philosophers' Imprint*.

Levinstein, B. A., & Herrmann, D. A. (2024). Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, 1-27.

Lewis, D. (1972). Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50: 249–258.

Li, M., Song, F., Yu, B., Yu, H., Li, Z., Huang, F., & Li, Y. (2023). Api-bank: A benchmark for tool-augmented llms. *arXiv preprint arXiv:2304.08244*.

Lin, E. (2021). The Experience Requirement on Well-Being. *Philosophical Studies* 178: 867–886.

Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J., & Stone, P. (2023). Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

Liu, R., Yang, R., Jia, C., Zhang, G., Zhou, D., Dai, A. M., Yang, D., and Vosoughi, S. (2023). Training Socially Aligned Language Models in Simulated Human Society. *arXiv preprint arXiv:2305.16960*.

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., ... & Tang, J. (2023). Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.

MacAskill, William ; Bykvist, Krister & Ord, Toby (2020). *Moral Uncertainty*. Oxford University Press.

Marcus, R. B. (1990). Some Revisionary Proposals about Belief and Believing. *Philosophy and Phenomenological Research* 50: 133–153.

Mill, J. S. (1863). *Utilitarianism*. Parker, Son, and Bourn.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice*. MIT Press.

Papineau, D. (1987). *Reality and Representation*. Blackwell.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint arXiv:2304.03442*.

Pettit, P. (2007). Responsibility Incorporated. *Ethics* 117: 171–201.

Place, U. T. (1956). Is Consciousness a Brain Process? *British Journal of Psychology* 47: 44–50.

Place, U. T. (2002). From Mystical Experience to Biological Consciousness: A Pilgrim's Progress? *Journal of Consciousness Studies* 9: 34–52.

Putnam, H. (1960). Minds and Machines. In Hook, S. (ed.), *Dimensions of Mind*, New York University Press. Reprinted in Putnam, H., (1975), *Mind, Language, and Reality*, Cambridge University Press, pp. 362–385.

Putnam, H. (1967). The Nature of Mental States. First published as "Psychological Predicates" in Capitan, W. H. and Merrill, D. D. (eds.), *Art, Mind, and Religion*, University of Pittsburgh Press. Reprinted in Putnam, H., (1975), *Mind, Language, and Reality*, Cambridge University Press, pp. 429–440.

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

Ryle, G. (1949). *The Concept of Mind*. The University of Chicago Press.

Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. *Noûs* 36: 249–275.

Schwitzgebel, E. (2011). Belief. In Bernecker, S. and Pritchard, D. (eds.) *The Routledge Companion to Epistemology*, Routledge, pp. 14–24.

Schwitzgebel, E. (2015). If Materialism is True, the United States is Probably Conscious. *Philosophical Studies* 172: 1697–1721.

Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*. Online First. <https://doi.org/10.1007/s43681-023-00379-1>.

Shaw, W. (1999). *Contemporary Ethics*. Wiley-Blackwell.

Singer, P. (1979) *Practical Ethics*. Cambridge University Press.

Smithies, D. (2019). *The Epistemic Role of Consciousness*. Oxford University Press.

Stalnaker, R. C. (1984). *Inquiry*. MIT Press.

Strawson, G. (1994). *Mental Reality*. MIT Press.

Sumner, W. (1996). *Welfare, Happiness, and Ethics*. Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N. Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All You Need. *arXiv preprint arXiv:1706.03762.*

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv preprint arXiv:2305.16291*.

Weatherson, Brian (2019). *Normative Externalism*. Oxford, UK: Oxford University Press.

Wiland, E (2022). What is Group Well-Being? *Journal of Ethics and Social Philosophy* 21: 1-23.

von Wright, G.H. (1963). *The Varieties of Goodness*. The Humanities Press.

Wu, Y., Prabhumoye, S., Min, S. Y., Bisk, Y., Salakhutdinov, R., Azaria, A., Mitchell, T, and Li, Y. (2023). SPRING: GPT-4 Out-performs RL Algorithms by Studying Papers and Reasoning. *arXiv preprint arXiv:2305.15486*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, *36*.

Yildirim, I., & Paul, L. A. (2023). From task structures to world models: what do LLMs know?. *Trends in Cognitive Sciences*.

Zhang, S., Zhang, J., Liu, J., Song, L., Wang, C., Krishna, R., & Wu, Q. (2024). Training Language Model Agents without Modifying Language Models. *arXiv preprint arXiv:2402.11359*.

Zhou, W., Jiang, Y. E., Cui, P., Wang, T., Xiao, Z., Hou, Y., ... & Sachan, M. (2023). Recurrentgpt: Interactive generation of (arbitrarily) long text. *arXiv preprint arXiv:2305.13304*.

Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., ... & Dai, J. (2023). Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.