Minimal epistemic agency in AI

Hadeel Naeem[1] & Julian Hauser[2]

Abstract

We argue that certain AI systems can be considered minimal epistemic agents — entities with the ability to achieve epistemic goals reliably. We approach epistemic agency through virtue reliabilism and identify two forms of minimal epistemic agency. The first requires counterfactual sensitivity to the reliability of one's belief-forming process: an agent should notice when their process becomes unreliable and cease forming beliefs with it. The second, adaptive minimal epistemic agency, adds an ability to maintain the reliability of one's process. Such an epistemic agent adapts to the world so that, rather than having to stop forming beliefs when their process turns unreliable, they can continue forming successful epistemic states. Applying this framework to AI systems, we show that current AI systems can exhibit counterfactual sensitivity to reliability. However, their limited capacity for self-modification constrains their ability to maintain reliability over time.

1. Introduction

We form many beliefs based on the output of AI systems. However, some philosophers argue that we are not justified in forming beliefs with AI output since these systems are opaque, do not model the world, do not have intentions, cannot testify, produce bullshit, or are simply stochastic parrots (Hicks, Humphries, and Slater 2024; Sparrow and Flenady 2025; Yildirim and Paul 2024; Bender et al. 2021). Philosophers who disagree with this sentiment struggle to explain why and when we are justified in forming beliefs based on AI outputs. In this paper, we want to weigh in on the side of this latter camp but propose to approach the topic from a specific angle: We believe that (certain) AI systems (in certain situations) can manifest an ability to reliably achieve epistemic goals. In other words, some AI systems are (at least minimal) epistemic agents.

If AI systems can be epistemic agents, then it would appear that we have at least some justification to form beliefs based on their outputs. This opens, among other things, the possibility of gaining knowledge based on an AI system's testimony, and potentially, them being scientific collaborators and epistemic authorities. In this paper, we hope to provide solid foundations for future work on responsibly relying on AI output.

[1]hadeel@hadeelnaeem.com

[2]julian@julianhauser.com

We approach the subject of epistemic agency with a virtue reliabilist framework. At the core of virtue reliabilism is the conviction that epistemic success (such as true belief, inquiry, knowledge, and so forth) is attributed to an agent for manifesting an ability. The focus on ability is essential – it ensures the agent's epistemic success is brought about by themselves and is non-accidental.

In section 2, we discuss how virtue reliabilism understands the minimal conditions for epistemic agency. The most minimal version of epistemic agency requires being counterfactually sensitive to the reliability of one's *belief-forming processes* (henceforth, CSRP).[3] According to this account, for someone to be an epistemic agent, or to responsibly employ their belief-forming processes, they must have an ability such that they would notice if their belief-forming process were to stop being reliable. Put another way, the agent should stop forming beliefs were circumstances to change in a way that rendered their processes problematic.

We think that CSRP captures the core of minimal epistemic agency, but we think it misses something important. In section 3, we highlight an additional aspect of (at least some forms of) minimal epistemic agency. Generally, epistemic agents, in addition to being CSRP, are able to maintain the reliability of their epistemic processes. They adapt their belief-forming processes to the changes they encounter in the dynamic world. Only with such an ability is the agent not left empty-handed when changes in the world render their belief-forming processes unreliable. This ability is essential for them to continue their epistemic quests. Call this – CSRP together with maintenance of reliability – *adaptive minimal epistemic agency*. We think this is an important part of (some forms of) minimal epistemic agency since epistemic agency isn't simply concerned with avoiding unsuccessful epistemic states but with *forming successful* epistemic states.

In section 4, we apply our adaptive minimal epistemic agency account to AI systems. Our discussions show that what current computer science calls *AI agents* do have some epistemic agency, in particular when it comes to being CSRP. We then go on to argue that because of their very limited capacity to change themselves, there are limitations regarding our second form of epistemic agency, namely, the maintenance of the reliability of the epistemic processes.

---

[3] Even though our understanding of epistemic agency is based on an account that looks at *beliefs*, we broaden this in section 4. Our account can therefore apply to AI systems even if it turns out that they do not instantiate beliefs.

2. Epistemic agency and counterfactual sensitivity to the reliability of the process

Virtue epistemologists, specifically virtue reliabilists, claim that knowledge is the product of cognitive ability (Sosa and Sosa 2013; Greco 2010; Pritchard 2010). More specifically, knowledge is cognitive success (true belief) we attribute to the agent for manifesting cognitive ability. To exercise such an ability, an agent must employ a reliable belief-forming process (one that forms far more true beliefs than false ones) such that the true belief is due to the agent manifesting relevant cognitive agency rather than some other factor (Pritchard 2010). In this paper, we call this relevant cognitive agency *epistemic* agency, and such agency is the subject of this paper.

Let's now turn to how epistemic agency (towards responsibly forming a belief) is manifested. To have such agency, one must have the ability to form and revise beliefs. An agent can manifest this ability in various ways. They can gather evidence regarding the reliability of their belief-forming processes and make decisions about what to believe. They can, when first learning to employ a new belief-forming process, make sure to employ the process in the right way and the right environment. All these actions manifest epistemic agency towards trying to get things right.

Many of the above forms of epistemic agency involve reflecting on one's belief-forming processes, but that's not always required. Virtue reliabilism is compatible with externalism about knowledge, and many virtue reliabilists contend that one can acquire knowledge even in the absence of reflective access to the reliability of one's belief-forming process. For instance, when we employ our perceptual faculties to form beliefs about what we see, we don't typically reflect on whether these faculties are reliable. On the virtue reliabilist account, even in the absence of reflective access to the reliability of one's belief-forming process, one can manifest epistemic agency sufficient for bringing about knowledge. Even in such cases, there is epistemic agency sufficient for knowledge. We will call this *minimal* epistemic agency and now go on to elaborate how an agent manifests such agency.

Put simply, minimal epistemic agency requires that the agent is counterfactually sensitive to the reliability of her belief-forming process (Greco 2010; Pritchard 2010; Palermos 2021; Naeem 2023). This means that the agent is in a relation to her belief-forming process such that if her process were to stop working reliably, the agent would notice a problem. We then say the agent is

counterfactually sensitive to the reliability of her process (CSRP). To be a minimal epistemic agent, the agent must be CSRP.

Since we manifest epistemic agency in the shape of being CSRP, we responsibly employ our perceptual faculties even when we do not reflect on them being reliable, Imagine the following situation: I go about my day seeing things and forming beliefs about what I see without reflecting on whether my perceptual faculties are working as they should. If suddenly everything started to turn red, I would notice that there is something wrong with my perceptual faculties, and I would not form the belief that the world has turned red.[4]

How an agent becomes CSRP is best understood in terms of Greco's account of cognitive integration. Cognitive integration is "the function of cooperation and interaction" between the belief-forming process and the other aspects of the agent's cognitive system (Greco 2010, 152). For a belief-forming process to integrate into the agent's cognitive system, the agent employs the process over a period of time. In this period, the agent forms various beliefs that cohere with existing beliefs and processes. When properly integrated, the agent's various preexisting cognitive processes can provide them with beliefs that can be compared to the belief-forming process's outputs. In this way, the agent can notice when the outputs of the process contradict their other beliefs, possibly indicating that the belief-forming process may have become unreliable.

The cognitive integration story doesn't end with integration on the level of beliefs. Integration also leads to a kind of fluency that allows a background monitoring of the belief-forming process (Palermos 2021; Shea et al. 2014). Fluency

[4] It's important to note that the account we describe only gives us a picture of *minimal* epistemic agency – and not epistemic agency more generally. This relates to the two grades of knowledge (Sosa 2001; J. Adam Carter 2024): minimal and reflective knowledge. Much of our reliance on our belief-forming processes is of the minimal kind. As mentioned, we routinely employ our perceptual and memory faculties without considering their reliability. Sometimes, we are involved in producing reflective knowledge that requires more of our epistemic agency. Here, we must reflect on the reliability of our processes (whether they are reliable and what makes them so, and so on). Ancient scepticists (Sextus 2006) teach us that chains of such reflective knowledge – where we have knowledge based on our knowledge about a belief-forming process's reliability – must end somewhere. According to the kind of account we embrace, they end with knowledge that issues from minimal epistemic agency.

describes how our belief-forming processes, when they are working well, are employed smoothly. When these processes become unreliable, this gives rise to metacognitive cues that indicate that something is amiss with the process. Here, the process's reliability isn't put into question because its outputs do not cohere with other beliefs, but because there are direct indications that the process is in a bad state.

Both integration with other beliefs and belief-forming processes and the establishment of metacognitive monitoring provide the agent with evidence when their belief-forming process no longer functions reliably. When integration is sufficient, the agent becomes CSRP: they would notice if their belief-forming process were to become unreliable.

An agent manifests minimal epistemic agency when they are CSRP. The agent is CSRP when they are in a position to notice if their process were to turn unreliable. If they are not in such a relation to their belief-forming process (that is, they would not notice were their process to turn unreliable), then the agent doesn't have sufficient agency towards responsibly forming beliefs or aiming at knowledge. In that case, even if the agent forms a true belief, this isn't their cognitive achievement, but rather dependent on environmental factors. When this happens, they aren't an epistemic *agent*. An epistemic agent should have an ability that significantly contributes to bringing about the success of their epistemic endeavours.

One important recent debate concerns the question of whether minimal epistemic agency can be even more minimal. (Pritchard 2010) suggests that the level of agency required to responsibly form a belief depends on the type of belief-forming process at hand and, importantly, when this process has been integrated. In the case of an innate biological process, we might responsibly form beliefs merely because the relevant processes have been employed over some period of time without giving rise to problems. In a sense, we rely on such innate faculties having proven their adequacy in the long game of evolutionary fitness. In contrast, when integrating a new tool-involving epistemic process, we might need to manifest more agency. Here, we might demand CSRP or even reflective knowledge that the tool involved is reliable.

This argument that different criteria for epistemic agency apply depending on the kind of belief-forming process involved opened the door for Clark (2015) to lower the bar even further. Clark is interested specifically in the question of extended belief-forming processes and argues against (what he takes to be) unjustified "bio-asymmetry" (J. Adam Carter 2022):Why demand more of

technologically extended processes than of bodily ones? Why not have the same minimal criteria apply across the board?

Clark (2015) suggests that epistemic agency can be realised by entirely subpersonal mechanisms that determine the reliability of the relevant epistemic processes. He works this out in terms of predictive processing, where certain mechanisms encode reliability weightings. The details do not matter for our purposes here. What does matter is that such epistemic agency only requires that subpersonal features of the cognitive system have a certain structure – it is not, in other words, required that the agent consciously notice when things are amiss. For a belief to be attributable to the agent, it is sufficient for there to be a subpersonal mechanism that ensures that the belief-forming process would no longer be employed were it to become unreliable.

One consequence of Clark's view is that immediate integration appears possible (Clark 2015; J. Adam Carter 2022). Since subpersonal conditions are sufficient for epistemic agency, the agent's involvement in epistemic agency may be extremely minimal. If epistemic agency can be achieved with such minimal involvement, it might be thought possible – and Clark explicitly endorses this possibility – for a belief-forming process to integrate immediately. To understand what is meant by this, it will help to contrast a case of typical integration with one of immediate integration. Consider someone who has learned to tell the time of day by looking at the position of the sun. They have acquired this ability through years of practice, developing a sense of how the sun's position maps onto the hour of the day and how these associations change throughout the seasons. Over the years, this process has been integrated into their cognitive system in various ways. For instance, this person also associates certain hours of the day with having specific levels of brightness, and this knowledge can be used to verify the proper functioning of the process. During integration, such knowledge of the brightness expected at certain hours of the day was also used to ascertain that the new process is reliable. Now, imagine a second agent who has the same ability to judge the time of the day. Unlike the first agent, this one was given the relevant ability by a magician who arranged his brain to match the first agent's. This second agent appears to be CSRP just as the first. According to Clark, we must in both cases judge the outputs of the agents' processes to be responsibly formed beliefs.

We have presented the account of minimal epistemic agency and the debates about even more minimal forms of epistemic agency and immediate integration. These debates have culminated in certain authors emphasising the

diachronic nature of integration and epistemic agency. It's to this that we now turn.

3. The diachronicity of epistemic agency

Most epistemologists have taken issue with Clark's claim that integration can be immediate. They instead argue that integration is a process that needs to happen over some period of time. We think this emphasis on the diachronic nature of epistemic agency is right, and we take it as an opportunity to discuss a second feature of epistemic agency that also springs from the diachronic nature of agency.

Disagreeing with the claim of immediate integration, virtue epistemologists such as Palermos (2021) and Pritchard (2023) have emphasised the importance of a gradual process of integration that takes place over some period of time in which the process is frequently employed by the agent. During this time, the beliefs formed with the help of this process must cohere with existing beliefs and give rise to yet more beliefs that assimilate into the agent's system. In other words, there must be a period of time during which the new process is integrated and where the agent's cognitive system coheres with the new processes (as described in section 2). This is how the new process gets appropriated by the agent and comes to be a proper part of them. The agent must not be bypassed, and when the agent is bypassed, we cannot speak of integration. Immediate integration is hence impossible.

We agree that immediate integration isn't possible, as in such cases, the agent doesn't manifest sufficient agency in acquiring the belief-forming process. Because the agent is bypassed, the resulting belief-forming process cannot be attributed to them.[5] We think this is one important regard in which epistemic agency cannot be made sense of without considering an agent's diachronic nature.

While we agree with these considerations concerning diachronicity, we do not think they go far enough. In particular, we think that epistemic agency requires not just becoming sensitive to a belief-forming process over some period of time but also being able to maintain the reliability of the belief-forming process across certain changes. CSRP is concerned with the agent's ability to notice if their belief-forming processes becomes unreliable in a changing world. But that's not all

[5] A more detailed critique of Clark's subpersonal epistemic agency account can be found in (Naeem 2023).

– an epistemic agent should also be able to maintain the reliability of their epistemic processes in a changing world. Epistemic agents exist across time, and our concept of epistemic agency should make sense of how they retain the viability of their epistemic processes as this happens.

Consider, again, the agent who has learned to tell the time of day by looking at the sun's position in the sky. Let's postulate that they have led a very sedentary life, never straying far from their place of birth. Now, imagine that one day they decide to start migrating south. As they move further and further away from their place of birth, their belief-forming process becomes less able to accurately inform them of the time of day. At a certain point, our agent notices this – their cognitive system flags the process as unreliable, and they consequently stop using the process to form beliefs about the time of day.

In a sense, and according to the existing literature on the topic, all is going well in this example. The agent doesn't form any false beliefs since they are CSRP. And because they are CSRP, the beliefs formed while their process is still reliable are formed responsibly. The agent manifests sufficient agency for the beliefs to be their beliefs and for them to potentially constitute their knowledge. However, this misses something important: the story ends when the agent stops employing their now unreliable belief-forming process. However, that isn't where epistemic agency ends. We do not only want to avoid forming false beliefs (and want the agent to be involved in avoiding them), we also want the agent to form true beliefs (and be involved in maintaining the relevant processes).[6]

The world we inhabit, both the environment and our own bodies, is highly dynamic. As the existing literature rightly acknowledges, when the world changes, our belief-forming processes may become unreliable, and epistemic agency requires that we are able to handle this. But an epistemic agent shouldn't just have at her disposal the option of throwing her hands up in the air and giving up the epistemic game. No, she should also have the ability to adapt her belief-forming processes so that they allow her to keep track of truths in the changing

[6] We do not, of course, demand that an epistemic agent should be able to appropriately handle all of the environmental changes that may interfere with their belief-forming processes. Some changes are such that the agent may not be able to steer their belief-forming process back to reliability. However, this is no different from the notion of being CSRP: for someone to be CSRP, they needn't be able to notice a malfunction in highly atypical situations (such as, in presence of knowledge undermining luck).

environment. As our agent travels south, we want them to adjust their belief-forming process so that they can now tell the time in different latitudes. We want to leave open how exactly this happens, making space for views that see the involvement of subpersonal processes as sufficient and views that require some reflexive engagement. What we insist on is that an epistemic agent should actively maintain the adequacy of their belief-forming processes and not just give up when things go wrong. Otherwise, epistemic agency is too brittle in a world that is as dynamic as ours.[7]

As already mentioned, the changes that necessitate maintenance work can stem from both the environment and our own bodies. Think, for instance, of the changes in our memory faculties as we grow older. When we notice that our memories have become less reliably true, we shouldn't just abandon forming beliefs based on them. Rather, we should strive to identify when our memory faculties are unreliable, and we should adjust our belief-forming process so that it is reliable again. For instance, we might not put as much of a stake in the truth of our beliefs about last year's events while still maintaining that memories from our younger years reliably convey what happened then.

We might worry that what we describe as maintenance of a belief-forming process's reliability is in fact already considered as part of the very concept of a reliable belief-forming process. After all, for a belief-forming process to be reliable, it needs to bring about true beliefs across a range of situations. Why demand that an agent additionally needs the ability to change their belief-forming process across a yet more extended range of situations? We think our demand is reasonable since the reliability of one's belief-forming process is one thing, and steering unreliable processes back to reliability is another. Consider again our

---

[7] Note that a similar condition can be applied to the agent being CSRP. Just as the agent should maintain her belief-forming processes in good order, she should also maintain her counterfactual sensitivity to their reliability. Imagine, again, our agent judging the time of day from the sun's position. A volcano erupts, darkening the sky at all hours. Since the agent uses the level of brightness to verify the outputs of her belief-forming process, she mistakenly concludes that her process has issued in the wrong verdict. Before, it made sense to say that a bright day should count against the reliability of her belief-forming process if that process issued in the belief that it's noon while it's dark outside. Now, this isn't correct anymore: it's now dark at noon. A good epistemic agent should now adjust her CSRP. Over time, she should no longer see darkness at noon as counting against the reliability of her belief-forming process.

agent, still in her village, forming beliefs about the time of day. Her belief-forming process is reliable, we submit, and issues in the formation of beliefs that are in the running for knowledge. Her belief-forming process would fail if she were to move to a significantly different location, but – for at least much of her life – this isn't a relevant possibility. When the agent moves southwards, some process does turn unreliable, and we want to say that it is the agent's belief-forming process that becomes unreliable. It's because this process turns unreliable that the agent needs to change herself to maintain its reliability.

To conclude, epistemic agency – at least of one central kind – involves not just a counterfactual sensitivity to the reliability of one's belief-forming process but an active maintenance of the reliability of one's belief-forming processes. We call this *adaptive* minimal epistemic agency. This form of epistemic agency highlights that we shouldn't just avoid forming false beliefs – we also want to form true beliefs.

4. AI, epistemic agency, and the importance of self-change

The virtue reliabilist concept of minimal epistemic agency is a useful framework to evaluate AI systems' epistemic agency. In this section, we use it to examine whether AI systems have CSRP and are able to maintain the reliability of their informational processes. In particular, we will look at large language models (LLMs) and the AI agents that are being built on top of them. We will evaluate where the current system's strengths and weaknesses lie and where AI design would need to be headed to become better epistemic agents.

Before we explore AI epistemic agency, some housekeeping is in order. First, some might worry that epistemic agency is a notion dependent on the concept of belief, and since AIs do not have beliefs, it follows that they cannot manifest epistemic agency. We think such arguments go a little too quickly and miss the core of our argument. The concept of minimal epistemic agency provides an understanding of when an agent minimally manifests an ability with which to non-accidentally achieve a successful epistemic state. While virtue reliabilist discussions prioritise true beliefs and knowledge, these aren't the only epistemic goals for which agents can manifest epistemic agency. And not all epistemic goals necessarily involve beliefs. We want there to be, for instance, space to attribute to certain animals the pursuit of epistemic goals without making this dependent on

their having beliefs.[8] Moreover, on the virtue reliabilist account, instead of beliefs, we could also talk about interrogative attitudes, and instead of knowledge, we could talk about the ability of aiming at sound questions (see J. Adam Carter and Willard-Kyle 2025). The main idea is that the core of virtue reliabilism isn't tied to forming beliefs or knowledge – the account can accommodate a diverse set of epistemic states. What is crucial for epistemic agency is manifesting an ability to produce an epistemic state so that the state's success is at least significantly creditable to their ability. We think that this concept of manifesting a relevant agency in bringing about a positive epistemic state is broad enough to apply to AI systems.

The question we care about, then, is whether AI can manifest epistemic agency in the pursuit of epistemic states broadly construed. For these epistemic states to be attributable to the system, it must be the case that the AI manifests an ability that makes it so that their epistemic states non-accidentally succeed. Otherwise, the relevant states might be successful, but, since the AI wasn't involved in their achievement in the right way, the epistemic success isn't attributable to it. It is then not an epistemic agent.

First, let's look at a bare LLM. As the reader is probably aware, such models are trained on large stocks of text data and then fine-tuned to be able to interact in conversations with human agents. The details may vary and are complex, but what is important for our purposes is that once such models have passed the learning stage and are available to users, they are fixed. In other words, they lack the ability to change themselves. It is, for instance, impossible for such a model to learn new things. While a user can interact with an LLM in such a way that the LLM has in its inputs – its *context window* – information about past runs, the LLM cannot do this without outside help. An LLM, on its own, doesn't have the ability to re-run itself. The consequent lack of memory has important consequences for their ability to adapt to a changing world – which is at the core of the notion of epistemic agency.

Because LLMs can't learn, they cannot be CSRP or maintain the reliability of their epistemic processes. To be CSRP, an agent must have some means of

[8] Alternatively, we could ascribe animals and AI agents something like a minimal belief (see Newen and Starzak 2022). Moreover, have argued that language agents can have beliefs (see Goldstein and Kirk-Giannini 2025), though others are skeptical (Herrmann and Levinstein 2024). One might read (Herrmann and Levinstein 2024) as saying that something like CSRP may be a good reason to attribute beliefs to AI.

detecting when one of their epistemic processes goes wrong, infer – when appropriate – that this is due to the process now being unreliable, and, finally, stop using the process. This chain fails already at the very first step: to detect that one's epistemic process has gone wrong, one must have some capacity to evaluate a past instantiation of one's epistemic process. This is impossible if the LLM never remembers anything about its previous operations. It follows that LLMs can't be epistemic agents. Similar considerations apply to our complementary notion of maintaining the reliability of one's epistemic process, but since LLMs already fail at the more basic task, we will not go into details here. [9]

AI agents, as understood by current computer science (see Zhang et al. 2025), are LLMs with scaffolds, where these scaffolds provide various additional functionalities that the AI can use by specifying so in its outputs and which can affect future inputs. This use of the term "AI agent" must not be taken to imply that such AI systems are agents in the philosophical sense we discuss. Of particular importance to us here is the ability to build up a store of information, which can then contribute to the context window in the LLM's future runs. Such an AI agent may, for instance, learn information about their user and then personalise its output to that user. Consider *ReadNow*, an AI agent designed to help a user discover interesting texts to read in their free time. The AI system does this by assessing the time spent by the reader on reading various topics and then identifying future reading opportunities. It has used this process to suggest readings for some time now and has been successful in recommending readings that interest its users. Currently, because the AI agent has observed the user spending considerable time reading about South Asia, and because it has information about the user's previous interests in economics, it has formed an informational state to the effect that the reader would like to read texts about Bangladesh's economy.

So far, the way that we have described this AI system, its goals, its function, and the steps it takes towards its goals, the AI system manifests no epistemic agency. It only employs certain processes that – if they go right – allow it to reliably

[9] Note that things are a little more complex regarding LLMs that can engage in *chain-of-thought* information processing. Such LLMs compute a chain of intermediate outputs, which build on each other and bring about the final outputs. For this reason, they can be said to have some limited form of memory about their past processing. Whether, and to what extend, they can be CSRP we leave open here, but the results are likely somewhere between the LLMs we discuss above and the AI agents to which we now turn.

achieve its epistemic goals. What we must consider now is whether the AI (a) is sensitive to the reliability of this process and (b) can maintain this reliability. Consider the following situation: The user doesn't take up any of the suggested readings about Bangladesh's economy but keeps reading avidly on topics in the field of economics and general news about South Asia. Of course, a single failure doesn't mean that the epistemic process as such has become unreliable, but with some more evidence, the AI agent confirms this. The AI agent's epistemic process that simply maps reading times to potential interests doesn't lead to reliable outputs anymore (maybe because the user is now forced to read on these topics for their work). If the AI system is to have epistemic agency, it must be able to detect when its epistemic process goes wrong in this way.

It appears possible that AI agents could, in fact, be CSRP. ReadNow's goal is to provide reading suggestions to its users, and it can use the user's uptake of its suggestions as independent verification of the reliability of its epistemic process. When very few reading suggestions are taken up, this may indicate that something is wrong with the way ReadNow forms the relevant informational states. When this happens, the AI agent can determine that the epistemic process is no longer fit to serve its purpose and stop employing it. To the extent that ReadNow is able to detect that the reliability of their epistemic process is failing in a range of counterfactual situations, we should say that it is a minimal epistemic agent.

The AI agent's epistemic agency is rooted in its ability to change itself: it can provision its memory scaffold with information about the aptness of its epistemic outputs and can use that information to determine whether its epistemic processes are in good order. Note, however, that an AI agent may only change its scaffold and that the LLM that forms the core of its information-processing system is off-limits. This entails that its ability to change itself is limited. For instance, it's not a given that an AI agent with an LLM fine-tuned to provide reading suggestions would be deterred from offering suggestions just because of an input that says its epistemic processes are no longer reliable. It's a well-known feature of AI agents that they ignore instructions when contrary ones are baked into their LLMs. However, we think that there are probably ways around such issues in many cases, so that AI agents can be CSRP at least sometimes. This means that AI agents have some minimal epistemic agency.

In the previous section, we argued that an epistemic agent shouldn't just manifest an ability to avoid false beliefs but that it should also have an ability to form true beliefs. Similar considerations apply to AI systems: we not only want them to avoid employing problematic epistemic processes, but also to employ

well-functioning epistemic processes. From a user perspective, we not only want AI systems to avoid giving us false information but also give us true information.

It is here, in the maintenance of the reliability of their epistemic processes, that AI agents are currently most limited. The problem has to do with the limited capacity for self-change we already touched upon: while AI agents can affect the content of their context window through changes in their scaffold, they cannot change the epistemic strategies encoded in their LLM. When these turn out to be inappropriate, they have no means of changing them other than specifying in the input that (a) the old process isn't reliable anymore and (b) describing the new process that should be followed. This is unlikely to work (reliably) for two reasons: first, since the problematic process is still encoded in the LLM, it may still affect the output even with contrary instructions in the input, and second, specifying alternative epistemic strategies of any complexity is likely difficult. It is far more difficult to specify a new or changed epistemic process than to avoid using a preexisting process. Except in rare cases, AI agents therefore cannot maintain the reliability of their belief-forming processes.

For AI agents to truly come into their own as epistemic agents, they need a stronger capacity to change themselves in the face of a dynamic world that might render their epistemic processes unreliable. AI agents must have the capacity to change their own epistemic processes when they detect that the processes are no longer reliable. In other words, must be able to maintain the reliability of their epistemic processes. The AI's epistemic processes must be, at least in some important regards, up to itself rather than being determined from the outside.[10]

[10] Interestingly, we have come to conclusions that link to the literature in the philosophy of artificial intelligence that deals with agency more generally. In some of this literature, an agent is seen, roughly speaking, as an entity that pursues goals through independent interaction with its environment. Those who have argued that certain AI systems can be such agents (Floridi and Sanders 2004; Nyholm 2018; Butlin 2024a, 2024b; Dung 2025) stress how certain AI systems may change themselves and modify the function associating inputs to outputs. We cannot then any more explain the system's behaviour by reference to some external influence – such as a designer's – but must rather refer to the system's endogenous ability to determine how to best achieve its goals. This is obviously only a very cursory glance at this already vast literature, but we hope it brings to the foreground interesting links for future research.

5. Conclusion

We looked at minimal epistemic agency on a virtue reliabilist framework that requires counterfactual sensitivity to the reliability of one's process. We then highlighted another important aspect of at least some forms of epistemic agency: maintenance of the reliability of one's epistemic process. We called epistemic agency that fulfils these conditions adaptive minimal epistemic agency. This is the epistemic agency of a system that adapts to epistemically relevant changes in the environment (rather than simply abandoning its now unreliable process).

We then examined whether AI technologies can exhibit minimal epistemic agency. Some AI systems, specifically AI agents, can manifest epistemic agency, especially on the traditional CSRP-based account. However, we think that AI agents fail to be adaptive minimal epistemic agents (at least in most cases). This is so because these systems have a very limited capacity for self-change. For adaptive minimal epistemic agency, AI agents must be able to change their belief-forming processes so that they remain reliable. This is difficult to achieve for an AI agent that cannot change significant parts of itself. Our work here should give AI engineers some clues on where improvements would most change things towards more epistemic agency. We should focus on how AI could better change itself in the relevant ways, which obviously has safety implications, and much further discussion here is necessary.

For now, we contend that some AI systems are at least minimal epistemic agents. If these systems meet the CSRP condition, we can understand their epistemic states as produced by their ability. By exhibiting such ability, these systems responsibly generate epistemic states. We can consider their informational states responsibly produced, at least in a minimal sense, because they exhibit an ability, where if their process were unreliable, they would not have produced the said information. If we are right, then some AI systems can exhibit minimal epistemic agency, and we have good reason to rely on their testimony.

References

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT '21. New York, NY, USA:

Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Butlin, Patrick. 2024a. “Reinforcement Learning and Artificial Agency.” *Mind & Language* 39 (1): 22–38. https://doi.org/10.1111/mila.12458.

———. 2024b. “The Agency in Language Agents.” *Inquiry* 0 (0): 1–21. https://doi.org/10.1080/0020174X.2024.2439995.

Carter, J Adam, and Christopher Willard-Kyle. 2025. “Virtue Epistemology for the Zetetic Turn.” *Mind* 134 (536): 943–66. https://doi.org/10.1093/mind/fzaf028.

Carter, J. Adam. 2022. *Autonomous Knowledge: Radical Enhancement, Autonomy, and the Future of Knowing*. 1st ed. Oxford University Press. https://doi.org/10.1093/oso/9780192846921.001.0001.

———. 2024. *Digital Knowledge: A Philosophical Investigation*. Routledge Studies in Epistemology. Abingdon, Oxon ; New York, NY: Routledge.

Clark, Andy. 2015. “What ‘Extended Me’ Knows.” *Synthese* 192 (11): 3757–75. https://doi.org/10.1007/s11229-015-0719-z.

Dung, Leonard. 2025. “Understanding Artificial Agency.” *The Philosophical Quarterly* 75 (2): 450–72. https://doi.org/10.1093/pq/pqae010.

Floridi, Luciano, and J. W. Sanders. 2004. “On the Morality of Artificial Agents.” *Minds and Machines* 14 (3): 349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Goldstein, Simon, and Cameron Domenico Kirk-Giannini. 2025. “AI Wellbeing.” *Asian Journal of Philosophy* 4 (1): 25. https://doi.org/10.1007/s44204-025-00246-2.

Greco, John. 2010. *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge ; New York: Cambridge University Press.

Herrmann, Daniel A., and Benjamin A. Levinstein. 2024. “Standards for Belief Representations in LLMs.” *Minds and Machines* 35 (1): 5. https://doi.org/10.1007/s11023-024-09709-6.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. “ChatGPT Is Bullshit.” *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Naeem, Hadeel. 2023. “Is a Subpersonal Virtue Epistemology Possible?” *Philosophical Explorations*, March, 1–18. https://doi.org/10.1080/13869795.2023.2183240.

Newen, Albert, and Tobias Starzak. 2022. “How to Ascribe Beliefs to Animals.” *Mind & Language* 37 (1): 3–21. https://doi.org/10.1111/mila.12302.

Nyholm, Sven. 2018. “Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci.” *Science and Engineering Ethics* 24 (4): 1201–19. https://doi.org/10.1007/s11948-017-9943-x.

Palermos, Spyridon Orestis. 2021. “System Reliabilism and Basic Beliefs: Defeasible, Undefeated and Likely to Be True.” *Synthese* 199 (3): 6733–59. https://doi.org/10.1007/s11229-021-03090-y.

Pritchard, Duncan. 2010. “Cognitive Ability and the Extended Cognition Thesis.” *Synthese* 175 (S1): 133–51. https://doi.org/10.1007/s11229-010-9738-y.

———. 2023. “Extended Knowledge and Autonomous Belief.” *Inquiry* 0 (0): 1–12. https://doi.org/10.1080/0020174X.2023.2238291.

Sextus. 2006. “Sextus Empiricus. 1: Outlines of Pyrrhonism.” In, Nachdr. The Loeb classical library 273. London: Heinemann.

Shea, Nicholas, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D. Frith. 2014. “Supra-Personal Cognitive Control and Metacognition.” *Trends in Cognitive Sciences* 18 (4): 186–93. https://doi.org/10.1016/j.tics.2014.01.006.

Sosa, Ernest. 2001. “Human Knowledge, Animal and Reflective.” *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 106 (3): 193–96. https://www.jstor.org/stable/4321199.

Sosa, Ernest, and Ernest Sosa. 2013. *A Virtue Epistemology*. Apt Belief and Reflective Knowledge / Ernest Sosa, Vol. 1. Oxford: Clarendon Press.

Sparrow, Robert, and Gene Flenady. 2025. “Bullshit Universities: The Future of Automated Education.” *AI & SOCIETY*, April. https://doi.org/10.1007/s00146-025-02340-8.

Yildirim, Ilker, and L. A. Paul. 2024. “Response to Goddu Et Al.: New Ways of Characterizing and Acquiring Knowledge.” *Trends in Cognitive Sciences* 28 (11): 965–66. https://doi.org/10.1016/j.tics.2024.08.004.

Zhang, Zeyu, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. “A Survey on the Memory Mechanism of Large Language Model-based Agents.” *ACM Trans. Inf. Syst.* 43 (6): 155:1–47. https://doi.org/10.1145/3748302.