Episteme (2026), 1–13
doi:10.1017/epi.2026.10129

episteme

ARTICLE

Artificial Competence

Ísak Andri Ólafsson

Faculty of Philosophy, History and Archaeology, University of Iceland, Reykjavík, Iceland
Email: isak.andri.olafsson@gmail.com

(Received 4 August 2025 UTC; revised 4 May 2026 UTC; accepted 25 June 2026 UTC)

Abstract

This article develops a virtue-epistemological distinction between general and narrow artificial intelligence (AI) systems by analysing AI competence through Ernest Sosa's virtue-theoretic framework. First, I outline a general distinction between narrow and general AI. I then characterise AI in a way that bridges its terminology with virtue epistemological concepts. Drawing on Sosa's idea that the telic structure is replicable for systems with merely functional-teleological aims, I argue that current AI systems can possess first-order competence. I argue that increasing domain-specific competence of narrow AI cannot elevate it to the status of general AI, as that requires the constitutional disposition to assess the aptness of one's own first-order performances, a second-order competence that no amount of first-order improvement can grant. Finally, I argue that the virtue-theoretic distinction between artificial narrow intelligence and artificial general intelligence depends on whether the AI system possesses second-order constitutional competence.

Keywords: Virtue epistemology; artificial intelligence; large language models; competence; Ernest Sosa

1. Introduction

Recent developments in generative artificial intelligence (AI) have forced us to reevaluate the difference between human cognition and machine intelligence, sparking critical questions about what requirements should be considered for generative AI systems, such as large language models (LLMs), to be thought of as possessing real intelligence.

This article attempts to answer this by introducing a virtue-theoretic epistemic explanation of AI competence, which allows us to have clear boundaries between artificial narrow intelligence (ANI) and artificial general intelligence (AGI), and furthermore provides the requirements AI would need to fulfil to be considered AGI. Ernest Sosa's virtue-theoretic framework is especially well suited to these questions because it treats competences as dispositions to perform well in a given domain, and because Sosa indicates that the telic structure is replicable for systems whose aims are

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

merely functional-teleological rather than epistemic in an agential sense.[1] This allows us to apply Sosa's framework to AI systems mutatis mutandis, without committing to the claim that current AI systems are epistemic agents.

2. Artificial intelligence

Recent innovations in both hardware and software have advanced AI technology dramatically in the last few years. The most important ones being a dramatic increase in processing power using specialised computational hardware (such as computer processing units, graphics processing units, tensor processing units,[2] and even early-stage qubit computing) (Zhu et al. 2023) and groundbreaking discoveries in machine learning using artificial neural networks (OpenAI 2023). This rapid development of AI has resulted in unprecedented breakthroughs in image processing and recognition (OpenAI 2024b), natural language processing (OpenAI 2023), audio manipulation (OpenAI 2024c), and autonomous systems more generally (OpenAI 2024d).

It will be useful moving forward to define narrow AI and general AI. Narrow AI commonly encompasses AI systems that are built for specific tasks or applications that are well defined. They execute precise functions within a limited domain that cannot be generalised to tasks beyond that domain, although general AI is commonly understood as having fully developed human cognitive capabilities (Sheikh Prins and Schrijvers 2023).

> **Artificial narrow intelligence**: Specialised AI systems engineered to perform well-defined tasks within a limited domain or a set of domains. These systems are often capable of surpassing human performance within the restricted domain, but incapable of generalising their performance beyond their designated functions (Goertzel and Pennachin 2007).

> **Artificial general intelligence**: AI systems with human-like cognitive abilities, which enable them to gather sensory inputs, reason, learn, and adapt across a diverse range of domains (Goertzel and Pennachin 2007).

ANI can be thought of as a task-specific optimised system in its relevant domain, whereas AGI can be said to display general cognitive autonomy as one would expect from normal epistemic agents (i.e. humans). AGI, as of writing this, does not exist. ANI, however, is rapidly integrating itself into our lives, and although image and audio generative AI systems have gotten increasingly more attention, LLMs, such as DeepSeek, Gemini, Claude, and ChatGPT,[3] have become almost synonymous with AI.

[1]Sosa (2021: 23) writes: 'the structure of our telic account is surely replicable for functional representations. [. . .] the aim might be just functionally teleological, as would then be the competence for its part'. Our claim, then, is not that any disposition of any reliable mechanism qualifies as a Sosa-style competence; that would collapse competence into causal reliability. The present framework requires that AI systems have an aim, given functionally by their design and training, and exercise SSS-structured dispositions to attain it. It does not require that they aim at truth in any robust agential sense. We thank an anonymous referee for pressing us to clarify this point.

[2]See, for example, Sato (2018) for further information on TPU processing.

[3]GPT stands for *generative pre-trained transformer*.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

LLMs are computational devices used for natural language processing that are able to generate human-like text and complete other language-related tasks with high accuracy (Kasneci Sessler Küchemann and Bannert 2023: 1). They are designed to generate sequences of words, code, or other data, from a source input (commonly referred to as a prompt) (Floridi and Chiriatti 2020). The most recent versions of LLMs at the time of writing, such as Anthropic Claude 5.x, Google Gemini 3.x, DeepSeek-V4, and ChatGPT-5.x, embody both the fantastical advancements of AI, as well as its inherent limitations. Although these advanced LLMs are able to instantly generate coherent and contextually relevant responses, easily passing the Turing test (Mei et al. 2024), they also generate perplexing non-sequiturs, tell brazen lies, and misunderstand the simplest of tasks (Hicks, Humphries, and Slater 2024).

Furthermore, even when LLMs produce coherent and accurate text, they are often unable to provide evidence for the claims they make. Instead, they make up imaginary sources and facts. This tendency of LLMs to fabricate evidence has been called 'AI hallucinations' (Hicks, Humphries, and Slater 2024: 38).[4] LLMs do not reflect on the text they produce, which can make these hallucinations (or confabulations) snowball,[5] generating further errors (Zhang et al. 2023: 1). Interestingly, in some cases, when LLMs are asked to justify their previous hallucinations, they generate false claims that they can recognise as being incorrect when they are presented with the same false claims in a separate interaction session (Zhang et al. 2023: 2). Recently, Hicks, Humphries, and Slater (2024) have argued that ChatGPT is a *bullshit machine*, and instead of hallucinations, ChatGPT's erroneous responses should be called *bullshit* because the terms 'hallucinations' and 'confabulations' suggest that ChatGPT can perceive (if it can hallucinate), or rely on, memory in a traditional sense (if it can confabulate) (Hicks, Humphries, and Slater 2024: 8). Furthermore, both terms suggest that ChatGPT is generally attempting to convey accurate information, when it is simply predicting the next word in a sentence.

Although we will be comparing AI attempts at a performance with human attempts to a degree, current LLMs only indirectly and incidentally track truths, as the design function of an LLM is to produce text that sounds plausible.[6] However, it is not clear what follows from this concession. Consider, for example, that the relevant design function of paradigmatic epistemic competences like perception and memory is to help survival and more generally evolutionary fitness of the organism. This is not an epistemic aim, but an evolutionary adaptive aim keyed to organism fitness. Granting this fact, however, does not stand in tension with the idea that our perceptual and memory functions reliably deliver accurate information.

The compatibility of epistemic faculties having non-epistemic design or etiological functions, thus non-epistemic function-generated aims, offers a vantage point to reassess what it means for LLMs to be optimised for producing plausible-sounding content. If that is their design function, they might have a function-generated aim that is not epistemic, but that in and of itself does not necessarily preclude them from having the kind of reliable connection with truth (a question determined by seeing how well they

[4]Edwards (2023) argues that, because there is a creative gap-filling principle at work without the intention to deceive, 'AI confabulations' would be a more accurate term.

[5]Zhang et al. (2023) refer to this phenomenon as *hallucination snowballing* (p. 1).

[6]ChatGPT, when prompted, replies that it adheres to guidelines to ensure that its responses are helpful, respectful, and safe. These guidelines are then further broken down into accuracy, relevance, respectfulness, safety, privacy, neutrality, transparency, ethical considerations, legal compliance, and empathy (OpenAI 2024a).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

actually deliver true information) that is demanded by competence, at least in appropriate conditions. For instance, if we specify that the kind of situational component of competence that is normal for LLMs to operate includes largely reliable training data, then it looks like we can make sense of a virtue-theoretic competence structure,[7] forming the basis of what looks like reliable artificial competence, even if we continue to grant that LLMs are intentionally designed to produce plausible human-like text.[8]

Note that the inner workings of AI can be akin to 'black boxes', so their justification for what led to a given prediction might not be interpretable by humans. Consider that there might be thousands of variables that contribute significantly to a single prediction. Even if they could all be inspected and the way the AI system decided to weigh each variable was accessible,[9] it is not reasonable to expect users to understand why the AI system came to a prediction (Ribeiro Singh and Guestrin 2016: 1137). To combat this, AI can be imbued with explicability,[10] which entails explainability and interpretability, often referred to as explainable AI, or explainable artificial intelligence (XAI). This enables humans to retain a level of intellectual oversight, generally by providing access to the reasoning behind the decisions and predictions made by the XAI, in the form of comprehensible explanations[11] that use concepts that we understand (Cappelen and Dever 2021: 25; Longo et al. 2024). Currently, many deployed AI systems (including LLMs) are not explainable, and no known methods exist to make them so without losing either information or accuracy (Rao 2025: 47). The opaque nature of AI systems makes XAI development focused on the explainability and interpretability[12] of AI vitally important to our understanding of present and future AI systems (Cappelen and Dever 2021: 26).[13] The literature on AI mirrors epistemology in some important ways. Sosa (2021) gives an example of an eye-exam, where one starts to incorporate guessing as the letters get smaller. In his case, the guesses are correct, and reliably so (Sosa 2021: 144). Guessing reliably has obvious parallels to LLMs. We will argue that, in both cases, these guesses are alethic affirmations, but they are not judgements, because the guesser aims at truth without aiming at aptness. The guesser's affirmation might be apt (correct because of manifested competence), but it is not aptly apt, because that would require aiming at aptness itself, guided by a second-order grasp that one's affirmation would be apt. This distinction between alethic affirmation and judgement will structure our account of the ANI/AGI divide.

[7]This is in reference to Sosa's SSS competence.

[8]Not everyone would characterise AI outputs in these terms. For example, Herman Cappelen and Josh Dever (2021), in their book *Making AI Intelligible*, have developed a de-anthropocentrised externalism that, along with the extended mind thesis, suggests that AI outputs can predicate properties.

[9]These kinds of AI systems have been called 'white boxes', and although they have an epistemological advantage over black-box models, 'they do not automatically yield solutions that facilitate responsible and accountable use in practice' because 'complex white-box models are also opaque to most domain- and non-experts' (Herzog 2022: 223).

[10]Explicability is a richer notion than explainability, as it combines intelligibility and accountability (Herzog 2022: 219; Floridi et al. 2018).

[11]Although, as Ribeiro notes, 'it is often impossible for an explanation to be completely faithful unless it is the complete description of the model itself', an explanation can still be meaningful if it is locally faithful, i.e. the explanation corresponds to how the model behaves in the proximity of the predicted instance (Ribeiro, Singh and Guestrin 2016: 1137).

[12]Christian Herzog has argued that there is a need to focus on explicability because neither explainability nor interpretability automatically incurs accountability (Herzog 2022: 223).

[13]XAI development is still plagued by implementation difficulties in extracting explanations of AI system behaviour, furthermore explicability and interpretability are not clearly defined, and it is not clear what kinds of tools are required for interpretability (Cappelen and Dever 2021: 26).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

Furthermore, the literature on AI is concerned with performances and domains, which fits well within a virtue epistemological framework such as the one Sosa has developed, most recently in his book *Epistemic Explanations* (2021). Additionally, Sosa's approach to competences takes them to be special cases of dispositions, i.e. dispositions of agents to perform well in a given domain. Thinking about competences as special cases of dispositions is advantageous to us while writing about AI, as machines can possess such dispositions in a functional-teleological sense, that is, dispositions to attain aims functionally provided by their design and training, which is independent of whether they possess consciousness or the kind of biological architecture on which human cognition supervenes. Before we consider AI competence from a virtue theoretical perspective, we need to elaborate on some of the key ideas found in Sosa's framework.

3. Virtue-theoretic competence

Sosa's telic virtue-epistemological view argues that epistemic normativity should be understood as one form of telic normativity and that performances can be understood as attempts aimed at some goal. An attempt can thus become an achievement when the attempt is successful because of sufficient competence (Sosa 2021: 18). As we have already established, competences are a special kind of dispositions, namely dispositions of agents to perform well in a given domain. To assess whether an agent's performance is good, Sosa proposes the following structure:

> **AAA (accuracy, adroitness, and aptness)**: Performances are Apt iff they are Accurate (they attain success) because they are Adroit, and they are adroit iff the performance issues from a complete competence (Sosa 2021: 18).

Aptness is a necessary and sufficient part of achievements, as without it one can be accurate and adroit because of luck, allowing for the possibility of lucky successes such as Gettier cases constituting cases of knowledge. An oft-cited example to grasp these distinctions is of an archer. The archer shoots at a target and their shot hits the target. The archer's performance is accurate because they hit the target. It is apt iff the archer hit the target because their performance was adroit, i.e. the archer manifested their competence when they hit the target.

An important question is what it means to manifest competence. In the archer example, the performance can be said to be completely competent if the shot manifested the archer's intrinsic archery skill (constitutional competence),[14] the archer was in good shape when shooting (inner competence), and the situation when shooting was appropriate (Sosa 2010: 465). Competences of an agent can thus be seen as dispositions they have to perform well and are comprised of the agent's skill, shape, and situation (Sosa 2010: 465).

> **SSS (seat, shape, and situation)**: Competence is manifested by the intrinsic skill situated within the agent (seat of the disposition), the agent's shape as it pertains to their current ability to exercise their skill (shape), and the appropriateness of the situation they are in with regard to how it affects their execution (situation) (Sosa 2010: 465).

[14]Sosa (2021) also refers to this as one's 'innermost competence' (p. 46).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

To further clarify, if the archer is to perform competently, they would first of all have to have the skills required to hit their target reliably enough when they are in proper shape and properly situated, as if they would not possess the necessary skills to do so they are incapable of competently hitting the target as luck would be the salient reason for them hitting it.

Second, they would have to be in a position to access those skills, i.e. the agent must be in shape. The details of what exactly constitutes good shape can vary in practice, but one can imagine that in the archer's case, being in shape includes being sober enough to utilise their skills, keep their eyes open, be mentally unperturbed, alert, and so on. Note that these requirements depend on the notion that without them the archer could not perform up to the intrinsic skill they possess, e.g. if they could reliably shoot their target while drunk with as much skill as if they were sober, then the state of being drunk would be compatible with the shape needed to be competent.

Finally, the situation must be appropriate for the archer to perform. Here, appropriate can be understood as specifying the range of conditions under which the archer's competence is properly attributable (i.e. conditions where their skill is meant to apply). In this specific case, the relevant range of situations might include, for example, wind levels and lighting conditions within the thresholds set by archery tournaments.

Now that we have explained what an apt performance entails, we must make a further distinction between apt performance, and aptly apt performance. Even when a performance is apt, in the sense that the agent is accurate because of adroitness, its aptness is not necessarily itself apt. For a performance to be aptly apt, the agent must be attempting not only to be accurate but apt. This becomes an especially meaningful distinction when we move from athletic performances to intellectual performances, where it maps onto the distinction between alethic affirmation and judgement.

When we apply the AAA structure to epistemic performances, we find certain parallels between the epistemic agent and the archer. On Sosa's (2021) view, the relevant epistemic performance is the alethic affirmation, an affirmation made in the endeavour to get it right on a given question (pp. 23–24). Animal knowledge can thus be understood within the AAA structure as an apt alethic affirmation: the agent affirms, the affirmation is accurate (true), adroit (issued from competence), and apt (true because of the manifested competence).[15] Once again, we use the SSS structure to see what competence in this example entails. First, the epistemic agent must possess the relevant constitutional epistemic skills, such as cognitive abilities (seat). Second, they must be in a good epistemic state to apply those skills, e.g. by being sober, awake, and alert (shape). Third, the situation must fall within a range that is not epistemically hostile, for example the epistemic agent is not being deceived by tricky lighting or illusions (situation).

To finalise our discussion of the virtue-theoretic framework that will be used to analyse AI competence, we need to categorise the kinds of knowledge that are derived from this framework. In Sosa's (2021) hierarchy, the lowest level of knowledge is animal

[15]Sosa equates animal knowledge with apt alethic affirmation (2021: 186), extending this even to sub-credal cases like the eye-exam subject (2021: 144). He also holds that the telic structure of his account is replicable for systems whose aims are merely functional-teleological (2021: 186). However, Sosa does not consider cases in which a system with merely functional-teleological aims produces apt alethic affirmations while plausibly lacking any form of mental life. LLMs may present such a case: they arguably satisfy the structural conditions for apt alethic affirmation through functional-teleological competence, yet whether this suffices for animal knowledge, even of the sub-credal variety, is not obvious. The present argument does not depend on resolving this question, as it only requires that ANI can produce apt alethic affirmations, not that they must constitute knowledge.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

knowledge, constituted by an apt alethic affirmation, one whose truth is due to the first-order competence of the thinker. This includes sub-credal cases like Sosa's eye-exam subject, who reliably gets the letters right at a row where they take themselves to be guessing (Sosa 2021: 144). Above this level is reflective knowledge full well, constituted by a fully apt judgement: an alethic affirmation whose aptness is itself attributable to the agent's meta-competence, that is a second-order sensitivity to when one's first-order affirmations would be apt (Sosa 2021: 186; Carter 2018: 285).[16]

4. Artificial competence

Sosa (2021: 23) extends the telic structure for evaluating performances to systems whose aims are merely functional-teleological, which includes ANI systems. Before turning to the more contested case of LLMs, however, it is worth constructing a clear view of ANI more generally. An AI chess engine's constitutional competence in the domain of chess is well defined and frequently exceeds the best human performers. Its situational requirements are tractable; a properly configured engine with sufficient computational power, given a legal position, manifests its disposition to evaluate the position and find the strongest move. The engine's aim here is thus not an epistemic aim directed at truth. Importantly, it has no capacity to step back and assess whether its evaluation function is inadequate for the position at hand or whether its opening book is being exploited. As the chess engine is rarely mistaken, it generally performs aptly within the domain, but it does not assess its own aptness.

If AI outputs can manifest the AI's competence to some degree, then it seems fruitful to examine AI competence further using the SSS model. Recall that SSS stands for seat, shape, and situation. In many ANI systems, we find that the seat, or constitutional competence, is not only in place, but oftentimes exceeds the highest performing humans in the relevant domain. What about ANI systems that are able to reach across domains? LLMs are capable of writing poetry and speeches, fluent in nearly every programming language, and have access to immense amounts of data, which they can quickly sort through to generate responses to prompts. One way to answer whether LLMs have constitutional competence is by examining whether the LLMs would be able to perform *reliably enough*[17] if they were in proper shape and in an appropriate situation (Sosa 2010: 473).[18]

So, let us examine LLMs' shape and situation and see how applicable they are to AI before we assess LLMs' constitutional competence. The shape, or inner competence, of ANI in general, including LLMs, seems unlike the inner competence humans possess because many of the examples generally used to describe this kind of competence, such as sobriety and alertness, do not apply to ANI. We can still imagine that a proper shape for an AI might consist of being configured correctly with access to sufficient

[16]Sosa (2021: 170–171, 186) identifies a third level above reflective knowledge full well: secure knowledge, which requires that the competences from which one's reflective knowledge derives must not be too easily lost while one retains the disposition to judge. Since the present article is concerned with the threshold between alethic affirmation and judgement, i.e. between animal and reflective knowledge full well, the conditions for secure knowledge are beyond the scope of this article. For a detailed overview of Sosa's framework, see Carter (2023).

[17]Reliably enough here is determined by the norms of the relevant domains of performance; for a detailed discussion on this, see Sosa (2010).

[18]This framing is inspired by Carter (2018).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

computational infrastructure to take advantage of its capabilities. This means the AI would not only have the appropriate nodes on the neural network after the relevant machine learning processes and the computational power to utilise them, but also the kind of configuration that enables it to correctly determine the appropriate response. To illustrate, think of an author that, when sober, can distinguish between fact and fiction, but when drunk they confuse the two and claim something as true that only happened in their book. In much the same way, LLMs have access to a vast amount of data, some of which is factual and some of which is fictional, and their shape determines whether they are configured to make proper use of the skills they possess.

Another way in which AI's shape could be compromised is when it gets stuck in a local minimum, unable to successfully perform. For clarification, think about a chess engine trying to find the best move. It sees the potential moves on the board and immediately dismisses the moves that do not seem promising. One such move is sacrificing the queen for apparently no compensation. As it turns out, that queen sacrifice is the best move at a very high depth. Unfortunately, the ANI is stuck with the candidate moves it initially decided on, trying to figure out which of them is most promising. In the case of LLMs, their structure allows them to access, transform, and employ, whatever data are needed. However, without a protocol to clearly distinguish between the two, it is not in proper shape to successfully perform in the epistemic domain of transmitting truths.

Regarding the range of appropriate situations in which an LLM's abilities are relevant, I suggest that it is defined by the quality and type of data used to train the LLM. If the situation is appropriate, i.e. the data fall within the type and quality the LLM's training is meant to accommodate, and the LLM is properly configured, it is clearly capable of performing well, with the caveat that their performance must be within their domain of expertise. In sum, ANI systems can possess complete competence if we define shape in terms of a system's ability to access and apply the skills it possesses, without having consciousness be a necessary property.

Even when LLMs reliably make true predictions within their domain of competence, the question of whether they are positioned to do so, whether their SSS are arranged to facilitate aptness, is itself an epistemic question. The requirements for reflective knowledge are inherently epistemic, regardless of whether the first-order competence is situated in an epistemic domain. A baseball player that is attempting to perform in the domain of baseball is not aiming at truth, but the epistemic domain is still relevant when they consider whether they are triple-S competent, as epistemic reflection is needed for this kind of a second-order competence. This does not mean that epistemic domains are uniform in nature, as standards between epistemic domains differ greatly depending on the setting (Sosa 2021: 14).

Such second-order competence is not present in LLMs either, even though their aims are epistemic to a degree. When an LLM generates a justification for its prediction, it is performing another first-order task. It is predicting what a good justification would look like given its training data, but it is not assessing whether the conditions of its own competence were such that its original performance was apt. Judgement, on Sosa's account, requires aiming at aptness (not merely at truth) (Sosa 2021: 24–25), along with a guiding presupposition that one's affirmation would be apt (Sosa 2021: 144–145). An LLM's evaluation of its own predictions lacks both, as it aims at truth on a new question (namely, 'was my prediction accurate?') without any presupposition concerning the aptness of the original affirmation. The second-order competence required for

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

judgement is not a separate performance but a condition on the original performance itself (Sosa 2021: 25, 186), and no amount of recursive self-monitoring can provide that.[19]

At a certain point ANI might become advanced enough that its predictions resemble those of Norman the clairvoyant, who in BonJour's (1980) thought experiment possesses a perfectly reliable clairvoyance faculty but lacks any reflective endorsement of that faculty. We would have little to no understanding of why the ANI makes a given prediction, but we could still be rather certain of its accuracy.[20] No matter how reliable such an ANI becomes, no increase in first-order competence can provide the second-order competence required for judgement. The difference between Norman and ANI is that although they are both making predictions without reflective endorsement, only the ANI, especially when equipped with high interpretability and transparency, can provide post hoc rationalisations supported by evidence and reasoning on account of the data it holds and the neural network processing it performs. However, this should not be mistaken for second-order epistemic reflection.

Recall the initial distinction between ANI and AGI and their defining characteristics, namely that ANI are specialised AI systems that perform tasks within a limited domain or a set of domains, whereas AGI have human-like cognitive capabilities across a diverse range of domains. Now, one might try to distinguish ANI from AGI by claiming that ANI performances are simply not apt, whereas AGI performances would be apt. But, as we have seen, this is too hasty. To illustrate, see the following example from Sosa (2021, 2010) about Simone the fighter pilot, who could easily not be in a real cockpit but in a near-perfect[21] simulation:

> In my thought experiment, trainees are strapped down asleep in their cockpits, and only then awakened. Let us suppose Simone to be in a real cockpit, flying a real plane, and shooting targets accurately. Surely her shots can then be not only accurate, but also competent, and even apt. (Sosa 2010: 468)

Simone's shots are apt: her shooting competence manifests in the accuracy of her shots. But, as Sosa points out, 'what of her intellectual shots, her judgments and beliefs?' (Sosa 2010: 468). Imagine that Simone forms the alethic affirmation that she successfully shot a target, and her affirmation is accurate and adroit. Sosa asks whether her alethic affirmation can also be apt, given that she could easily have been placed into a simulation (Sosa 2021: 169–170; Sosa 2010: 468). It is difficult to deny that it can be, given how clearly apt her physical shot is despite the fragility of her situation (Sosa 2010: 468). Simone's 'competence manifests in the accuracy of her shot', and the same holds for her epistemic performance (Sosa 2010: 468). On Sosa's recent view (2021: 169–170), Simone can even attain reflective knowledge (but not secure knowledge).[22] Nevertheless, the case

[19]I am grateful to an anonymous referee for prompting further clarification on this point.

[20]Process reliabilists are thus faced with a new version of the clairvoyance problem in the form of a black-box ANI whose alethic affirmations are apt without any second-order grasp of why those affirmations are apt.

[21]Imagine the simulation to be identical to the real world except for whatever property distinguishes the real world from a simulation, i.e. there are 'no tell-tale signs' that Simone is in a simulation (Sosa 2010: 468).

[22]On Sosa's earlier view (2010: 474–475), Simone's case was explained by distinguishing animal from reflective knowledge: Simone possesses the former but lacks the latter, since she cannot discern whether her first-order competence obtains and thus lacks the situational component of second-order competence. On Sosa's more recent view (2021: 169–171), once default assumptions are admitted into the framework,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

shows how one can possess first-order competence while lacking the situational security that higher grades of knowledge demand. We can imagine ANI systems existing in a kind of simulation, trained on datasets that cannot fully represent the world, leaving them incapable of recognising the incompleteness of their information.[23] ANI systems, unlike Simone, cannot reach the level of reflective knowledge, because they lack the second-order competence required for judgement.[24]

Sosa divides the various epistemic competences into two main categories: non-global competences, which host seemings, and global epistemic competences, which are basic judgement-forming competences (Sosa 2021: 148).[25] Global epistemic competences differ from non-global competences in the sense that they are used to determine how to judge, *all things considered*.[26] This means that nearly anything could be relevant to any epistemic attempt (given proper stage-setting). Sosa claims that this kind of 'holistic competence is global in that it is required in properly making *any* judgment or forming *any* belief' (Sosa 2021: 148).

To demonstrate the difference between global and non-global competences, Sosa gives an example of Norm, the normal perceiver, and Abnor, the mental ward patient (Sosa 2021: 148).[27] On any given day when Abnor wakes up, he could have been experimented on which would deprive him of some, or all, epistemic competences that day (both non-global and global). If he would wake up without non-global epistemic competence, then he would not be capable of reflective knowledge unless someone would tell him that his competence, be it his shape or situation, was absent (and he would be in a position to make proper use of that information) (Sosa 2021: 151). If he would wake up without global epistemic competence, then there is no way for him to properly reflect on any tell-tale signs that this has happened, and he would lack the

Simone can assume by default that she is in a genuine cockpit, and thereby attain reflective knowledge as well; what the simulation scenario defeats is not her reflective knowledge but the security of that knowledge, since her relevant competence could too easily have been absent (Sosa 2021: 170–171). The present article's use of the Simone case does not depend on the pre-revision diagnosis.

[23]One objection is that LLMs often seem aware of their own limitations. For example, when an earlier version of ChatGPT, which did not have access to up-to-date information from the internet, was asked about events occurring after January 2022, it provided the following disclaimer: 'My training data includes information up until January 2022. Anything that has occurred or been published after that date wouldn't be directly accessible to me unless it has been shared with me in this conversation' (OpenAI 2024a). However, this response only reflects information provided by those who control its training data. To illustrate, if Simone were told that she was in a simulation, she might form the corresponding belief, but that would not enable her to determine whether her shooting would remain accurate outside of the simulation. Similarly, LLMs that have been informed of their limited datasets are not in a better position to account for their limitations.

[24]Simone's predicament arises at a higher threshold between reflective and secure knowledge, which is not relevant to the ANI/AGI distinction as we define it.

[25]This distinction between global and non-global competences is reminiscent of a distinction that can be found when considering different kinds of XAI explanations, which can be categorised as either being global or local. Global explanations are best understood as being two-place explanations, whereas local explanations are three-place explanations. That is, global explanations can be generally defined as explaining AI outputs, whereas local explanations can be defined as explaining specific outputs of an AI.

[26]Sosa (2021) finds that whether such global competences exist is an empirical question (p. 148).

[27]I am using Sosa's example for a different purpose than the one it serves in his discussion. Although he introduces the scenario to show the difference between the *threat of lacking* either global or non-global epistemic competence, my focus is on the distinction between the two if that threat is *realised*.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

ability to assess its presence (Sosa 2021: 149, 151).[28] According to Sosa, Norm is positioned to acquire both animal and reflective knowledge, whereas Abnor is limited to animal knowledge (Sosa 2021: 151).

For an AI to possess the sort of holistic global epistemic competence, Sosa identifies, it would need to be able to make judgements *all things considered*. One might think that Norm and Abnor show ANI could potentially attain reflective knowledge if its architects were to inform it that its competence had been compromised. But this would not amount to reflective knowledge. Even if the ANI were so informed, it would still lack the global epistemic competence (which entails second-order competence) that would be required to integrate that information into a revision of its own epistemic standing. And in any case, the architects would be the ones reflecting, not the ANI itself.

When it comes to the loss of global competence, we see that ANI is in the same or worse position as Abnor. At best it can only tell when its global epistemic competence has not been compromised with no way of knowing when it has, at worst it cannot even tell when it has not been compromised, because it lacks the second-order competence to assess the status of its own global competence (Sosa 2021: 152). For an example, consider that ANI systems can hallucinate while being unaware they are doing so because they lack second-order competence. In that case, adding data does not make up for this deficit, as ANI cannot verify that data to be correct, and is incapable of reflecting on that unreliability. ANI's deficit is therefore constitutional; it lacks second-order competence not because of an unfavourable environment, but because its seat does not include the capacity to assess its own first-order competence.

I propose the following virtue-theoretic definitions of ANI and AGI from competence:

> **Virtue-theoretic ANI**: An AI that can possess complete first-order competence within some domain, whose predictions are alethic affirmations that can be apt, but which lacks the second-order competence that would make those affirmations judgements.[29]

> **Virtue-theoretic AGI**: An AI whose performances are judgements[30], or alethic affirmations aimed at aptness, in virtue of possessing, as part of its constitutional competence (seat),[31] the second-order competence to assess whether its own first-order performances would be apt.

[28]Although Sosa (2021) points out that if it is not missing then it is possible to know that it is not missing in virtue of having the necessary competence to know so (p. 152).

[29]Whether apt alethic affirmations by systems with merely functional-teleological aims constitute sub-credal animal knowledge is a further question that is beyond the scope of the present argument. What matters for the ANI/AGI distinction is that no such affirmation, however apt, rises to the level of judgement.

[30]One might object that judgement and reflective knowledge require conscious reflection, which would make these attributions to AGI incoherent without arguing that they possess consciousness. Sosa addresses this concern directly: 'We here abstract from that to a reflection that needs to be mental but need not be conscious. This is in line with the now recognized subconscious depths of the mind. Our proposed abstraction makes room for a subconscious "reflection" constituted by mental (though not necessarily conscious) reference of the mind to itself' (Sosa 2021: 186). Whether AGI would possess such a non-conscious mental capacity is, however, beyond the scope of this article.

[31]The criteria the framework imposes on a system's seat are threefold: the system must aim at aptness rather than truth alone (Sosa 2021: 24–25), possess a guiding presupposition that its first-order affirmations would be apt (Sosa 2021: 145), and this second-order state must constitute mental self-reference (Sosa 2021:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

5. Conclusion

Having a clear virtue-theoretic distinction between ANI and AGI enables us to assess AI systems using an established and robust virtue epistemological framework. We can see how ANI, while capable of apt alethic affirmations, cannot make judgements regardless of its level of first-order competence because it lacks the required second-order competence. ANI is therefore not simply a less developed AGI; the gap between them is qualitative rather than quantitative, since no improvement in first-order reliability can produce the constitutional disposition to assess the aptness of one's own performances. For a system to qualify as AGI, it must possess second-order constitutional competence, although note that actual reflection on any given occasion is not required. An agent placed in a fake barn case possesses human-like cognition even when unable to aptly reflect on a particular alethic affirmation. The constitutional capacity to reflect, not its successful exercise in every instance, is what separates AGI from ANI. A surprising result here is that AGI, on the current account, runs into many of the same problems epistemologists have been working on, such as the problem of epistemic luck.[32]

Financial support. The author declares that no funds, grants, or other support were received during the preparation of this article.

Competing interests. The author declares none.

Author contributions. This manuscript was written in full by the author, Ísak Andri Ólafsson.

Ethical standards. Not applicable.

Consent to participate. Not applicable.

Consent to publish. Not applicable.

References

BonJour L. (1980). 'Externalist Theories of Empirical Knowledge.' *Midwest Studies in Philosophy* **5**, 53–73. https://doi.org/10.1111/j.1475-4975.1980.tb00396.x

Cappelen H., and Dever J. (2021). *Making AI Intelligible: Philosophical Foundations*. New York: Oxford University Press. https://doi.org/10.1093/oso/9780192894724.001.0001

Carter J.A. (2018). 'Virtue Epistemology, Enhancement and Control.' *Metaphilosophy* **49**, 283–304. https://doi.org/10.1111/meta.12304

Carter J.A. (2023). *Stratified Virtue Epistemology: A Defence*. Cambridge: Cambridge University Press.

Edwards B. (2023, April 6). *Why ChatGPT and bing chat are so good at making things up*. Retrieved from Ars Technica: https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/

Floridi L., and Chiriatti M. (2020). 'GPT-3: Its Nature, Scope, Limits, and Consequences.' *Minds and Machines* **30**(4), 681–94. https://doi.org/10.1007/s11023-020-09548-1

Floridi L., Cowls J., Beltrametti M., Chatila R., Chazerand P., Dignum V., Luetge C., Madelin R., Pagallo U., Rossi F., Schafer B., Valcke P., and Vayena E. (2018). 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.' *Minds and Machines* **28**, 689–707. https://doi.org/10.1007/s11023-018-9482-5

186). The conditions under which a system could satisfy the third criterion are contested; this is arguably where the most demanding philosophical work on AGI, on the present account, lies. We thank an anonymous referee for prompting this clarification.

[32]I thank Adam Carter, Rory Aird, and an anonymous referee for their helpful comments and advice. This work was supported by the Royal Institute of Philosophy.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129

Goertzel B., and Pennachin C. (2007). Contemporary Approaches to Artificial General Intelligence. In B. Goertzel, and C. Pennachin (eds), *Artificial General Intelligence*, pp. 1–28. Heidelberg: Springer Berlin. https://doi.org/10.1007/978-3-540-68677-4

Herzog C. (2022). 'On the Risk of Confusing Interpretability with Explicability.' *AI and Ethics* **2**, 219–25. https://doi.org/10.1007/s43681-021-00121-9

Hicks M., Humphries J., and Slater J. (2024). 'ChatGPT is Bullshit.' *Ethics and Information Technology* **26**(38), 1–10. https://doi.org/10.1007/s10676-024-09775-5

Kasneci E., Sessler K., Küchemann S., and Bannert M. (2023). 'ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education.' *Learning and Individual Differences* **103**, 1–9. https://doi.org/10.1016/j.lindif.2023.102274

Longo L., Brcic M., Cabitza F., Choi J., Confalonieri R., Del Ser J., Guidotti R., Hayashi Y., Herrera F., Holzinger A., Jiang R., Khosravi H., Lecue F., Malgieri G., Páez A., Samek W., Schneider J., Speith T., and Stumpf S. (2024). 'Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions.' *Information Fusion* **106**, 1–22. https://doi.org/10.1016/j.inffus.2024.102301

Mei Q., Xie Y., Yuan W., and Jackson, M.O. (2024). 'A Turing Test of Whether AI Chatbots are Behaviorally Similar to Humans.' *Proceedings of the National Academy of Sciences* **121**(9), 1–8. https://doi.org/10.1073/pnas.2313925121

OpenAI. (2023). GPT-4 Technical Report. Retrieved from https://cdn.openai.com/papers/gpt-4.pdf

OpenAI. (2024a). ChatGPT. Retrieved from https://chat.openai.com/chat

OpenAI. (2024b, February 15). *Video generation models as world simulators*. Retrieved from OpenAI: https://openai.com/index/video-generation-models-as-world-simulators/

OpenAI. (2024c, June 7). *Expanding on how Voice Engine works and our safety research*. Retrieved from OpenAI: https://openai.com/index/expanding-on-how-voice-engine-works-and-our-safety-research/

OpenAI. (2024d, May 13). *Hello GPT-4o*. Retrieved from OpenAI: https://openai.com/index/hello-gpt-4o/

Rao S. (2025). The Limits of AI Explainability: An Algorithmic Information Theory Approach. *arXiv:2504.20676*. Retrieved from https://arxiv.org/abs/2504.20676

Ribeiro M.T., Singh S., and Guestrin C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). San Francisco: Association for Computing Machinery. https://doi.org/10.18653/v1/n16-3020

Sato K. (2018, August 30). *What Makes TPUs Fine-Tuned for Deep Learning?* Retrieved from Google Cloud Blog. https://cloud.google.com/blog/products/ai-machine-learning/what-makes-tpus-fine-tuned-for-deep-learning

Sheikh H., Prins C., and Schrijvers E. (2023). 'Artificial Intelligence: Definition and Background.' In H. Sheikh, C. Prins, and E. Schrijvers (eds), *Mission AI*, pp. 15–41. Cham: Springer. https://doi.org/10.1007/978-3-031-21448-6_2

Sosa E. (2010). 'How Competence Matters in Epistemology.' *Philosophical Perspectives*, **24**, 465–75. https://doi.org/10.1111/j.1520-8583.2010.00200.x

Sosa E. (2021). *Epistemic Explanations: A Theory of Telic Normativity, and What It Explains*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198856467.001.0001

Zhang M., Press O., Merrill W., Liu A., and Smith N.A. (2023). How Language Model Hallucinations Can Snowball. *ArXiv preprint, 2305, 13534v1*. https://doi.org/10.48550/arXiv.2305.13534

Zhu S., Yu T., Xu T., Chen H., Dustdar S., Gigan S., Gunduz D., Hossain E., Jin Y., Lin F., Liu B., Wan Z., Zhang J., Zhao Z., Zhu W., Chen Z., Durrani T. S., Wang H., Wu J., Zhang T., and Pan Y. (2023). 'Intelligent Computing: The Latest Advances, Challenges, and Future.' *Intelligent Computing* **2**(0006), 1–45. https://doi.org/10.34133/icomputing.0006

Ísak Andri Ólafsson is a member of the Institute of Philosophy at the University of Iceland. He works primarily in epistemology, with a focus on social epistemology. Further information can be found at www.isakandriolafsson.com, and he can be reached at isak.andri.olafsson@gmail.com.

Cite this article: Ólafsson Í. A. (2026). "Artificial Competence." *Episteme* 1–13. https://doi.org/10.1017/epi.2026.10129

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 09:26:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10129