Counter-closure principles in the age of complex software systems: a generalized challenge from AI

Matteo Baggio

matteo.baggio@unito.it

Università degli Studi di Torino

Forthcoming in *Social Epistemology* – pre-proof draft

Do not cite without permission

Abstract: The rapid advancement of artificial intelligence has brought a host of new epistemological challenges. One particularly pressing question is whether, and to what extent, AI systems can serve as sources of epistemic goods. Can they effectively transmit knowledge or understanding? And if they do not possess these epistemic goods themselves, can they still generate them for human users? This article explores these questions by critically examining the constraints posed by counter-closure principles – epistemological principles that allegedly cast doubt on the epistemic potential of AI. By addressing these principles, we aim to lay the groundwork for a systematic inquiry into the social epistemology of AI.

Keywords: artificial intelligence; counter-closure principles; propositional knowledge; epistemic warrant; phenomenal understanding; epistemic transmission; testimony.

1. Introduction

The rise of artificial intelligence has introduced a range of complex epistemological challenges. One of the most pressing issues is whether – and to what extent – AI systems can serve as sources of epistemic goods, such as knowledge and phenomenal understanding, for their users.[1] This question becomes especially urgent given that we are not yet in a position to attribute the corresponding doxastic (or noetic) states to the AI systems themselves.

In this article, we address these issues by critically examining the constraints imposed by counter-closure (*CC*) principles – epistemological principles that purportedly cast doubt on the capacity of AI to produce genuine epistemic goods for its users. Our aim is to lay the groundwork for a systematic exploration of AI's epistemological role.[2]

Before proceeding, it is essential to clarify a central assumption of this paper: specifically, the importance of focusing on *CC* principles. There are two main reasons for this focus. First, as we will soon argue, the logic underlying *CC* principles appears to govern the transmission of a broad range of epistemic goods from traditional epistemic sources, such as inference and testimony. Hence, if we are to evaluate AI systems by the standards applied to traditional goods and sources, it is reasonable to expect that *CC* principles should apply to AI as well. Second, suppose *CC* principles were to fail in accurately capturing the necessary conditions for acquiring traditional epistemic goods. In that case, any adequate account of these principles' failings should explain why such conditions are not required, neither for humans nor for AI systems. In light of this, determining whether *CC* principles truly reflect the necessary conditions for fostering the acquisition of epistemic goods is essential for understanding the broader epistemological role of AI.

[1] Understanding comes in different kinds and can be directed to different objects. In this paper, we focus on phenomenal understanding, i.e., the kind of understanding that is directed to phenomena of reality. There is significant amount of disagreement in the literature about what phenomenal understanding involves. The literature is divided into two broad camps: *explanationists* and *non-explanationists*. Explanationists take understanding to be reducible to knowledge of an explanation (see for instance Khalifa 2017). Non-explanationists, in contrast, deny this; understanding does not entail knowledge of explanations (see for instance Greco 2014; Grimm 2014; Bengson 2015). In this paper, we do not aim at settling the issues over the nature of phenomenal understanding. However, for the sake of our analysis, we will assume that phenomenal understanding is to be spelled out along non-explanationists lines. The reason for this choice is twofold. It seems to us that explanationists accounts demand at the same time too much and too little for understanding. They demand too much, because understanding something seems to succeed without (explicit) knowledge of an explanation (cf. Lipton 2009; Kelp 2015; Elgin 2017). They also demand too little, because if knowledge of an explanation is all that is required for understanding, a sufficiently reliable software able to deliver highly complex and barely understandable explanations would work as a source of phenomenal understanding for the users who believe its outputs (cf. Malfatti 2025). This would set the bar for phenomenal understanding too low.

[2] In what follows, the term "AI" will refer to computer systems or software designed to perform tasks that typically require human intelligence, such as problem-solving, language understanding, and decision-making. Most existing AIs are narrow AIs, designed for specific functions such as chatbots or image recognition. Unlike speculative forms of AI, such as general AI (human-level cognition) or super AI (beyond human intelligence), narrow AI operates within defined limits but can adapt over time through methods like machine learning and natural language processing. These systems learn from data sets, refining their performance through iterative processing. As Grodzinsky, Miller, and Wolf (2020) note, while basic forms of narrow AIs behave predictably, more advanced ones adjust based on new inputs, enabling them to respond dynamically. In this paper, "AI" refers to such adaptive, task-specific systems. That said, we acknowledge that the term "AI" is a subject of contention in contemporary scholarship. Some scholars argue that the label is misleading, as current AI systems may lack genuine intelligence or understanding (see, for instance, Bender et al. 2021; Floridi 2023), preferring terms such as "machine learning systems", "algorithmic tools", or "computational models". However, in what follows, we will retain the conventional term "AI" for several reasons. First, it is standard in both academic literature and public discourse to examine the systems we will consider. Second, our epistemological analysis focuses on the capacity of these systems to process information and generate outputs that users treat as testimony-like, regardless of whether they possess genuine intelligence. Third, since intelligence itself is a contested concept, any terminological choice remains potentially controversial. In this sense, then, readers should understand our use of "AI" as referring to the characteristics described above without strong claims about the underlying nature of intelligence in these systems. Indeed, as we will soon see, our central thesis – that epistemic goods can be generated even by sources lacking corresponding mental states – is strengthened by acknowledging that current AI systems may lack genuine intelligence. Thanks to an anonymous referee for requesting greater clarity on this important point.

The article is structured as follows. In §2, we introduce *CC* principles and explain how they establish the necessary conditions for transmitting epistemic goods through inference and testimony. In this section, we also discuss how this family of principles can be expanded to include a wider range of epistemic goods. In §3, we briefly explore the limitations of the view that inference always transmits epistemic goods. Finally, in §4, we turn to testimony, engaging with Lackey's well-known cases and reinterpreting them through the lens of new epistemic goods and emerging epistemic technologies.

2. Counter-closure

Traditionally, *CC* principles in epistemology have been examined for their role in specifying the conditions under which inferential epistemic goods, such as knowledge and epistemic warrant, can be acquired. A central example of this family of principles is the *CC* principle concerning inferential knowledge (*CCK*), which can be stated as follows:

> *CCK* Necessarily, if (i) an epistemic agent *S* believes *q* solely on the basis of a competent inference from *p*, and (ii) *S* knows *q*, then *S* knows *p*.

As Luzzi (2019) aptly observes, *CCK* is widely regarded as an intuitive and uncontroversial principle of inferential knowledge. This consensus is evident in the way leading epistemologists discuss it. For example, Williamson (2007, 145) and Audi (2010, 184–5) both endorse *CCK* without offering explicit arguments in its defense, and Williamson (1994, 222) even appeals to it directly in support of his margin-for-error principle within his account of vagueness.

Yet, despite its intuitive appeal, two important qualifications about *CCK* should be noted. First, condition (i) of *CCK* excludes cases of *epistemic overdetermination*, meaning that *CCK* applies only in cases where the sole epistemic source for an agent's belief that *q* is the inference from *p*. Second, *CCK* operates as the inverse of traditional *closure* principles for knowledge, meaning that while closure principles outline how knowledge extends from premises to conclusions in deductive inferences, *CCK* works in reverse, tracing knowledge backward, from conclusions to premises.[3]

Now, as previously mentioned, *CC* principles are also highly adaptable and can be reformulated to apply to epistemic goods beyond knowledge. For instance, by substituting "warrant" for "knowledge" in *CCK*, one can derive plausible variants of *CC* principles applicable to both *factive* and *non-factive* epistemic goods (cf. Murphy 2015; Baggio 2025). This adaptability is philosophically significant. If, as many argue, some epistemic concepts reduce to others, then *CC* principles – by laying out necessary conditions for both factive and non-factive goods – play a crucial role in ensuring that these concepts remain governed by principled constraints (cf. Schechter 2017). From this broader perspective, then, the family of *CC* principles not only holds intuitive appeal but also occupies a central and philosophically important place in the structure of epistemological theorizing.[4]

As further evidence of their reach, *CC* principles are not limited to inferential sources. They can also be extended to other epistemic contexts, such as testimony. Consider this variant:

> *CCT* Necessarily, if (i) a receiver *R* believes that *p* solely on the basis of a sender *S*'s testimony that *p* and (ii) *R* knows that *p*, then *S* knows that *p*.

[3] To get a version of the closure principle for knowledge, the reader can simply substitute *p* with *q* and *q* with *p* in (ii) and the consequent of *CCK*, respectively.

[4] For example, when it comes to phenomenal understanding, this feature of the *CC* principles helps maintain a balanced neutrality between both reductionist and non-reductionist perspectives.

Like *CCK*, *CCT* is intuitively compelling and widely endorsed (cf. Burge 1993, 486; Audi 1997, 410). Moreover, both principles capture the thought that inference and testimony primarily function as vehicles for transmitting epistemic goods. From this perspective, the success of transmission depends on such goods already being present at the source.[5]

To illustrate, consider the case of inference first. Suppose *S* believes *q* solely through a competent inference from *p*. If *p* were to lack any epistemic standing – if it weren't known or warranted by *S* – then there would be nothing to transmit to *q*. Hence, for *S* to know *q* based solely on *p*, *S* must also know *p*. The same logic applies to testimony. If *R* believes *p* based exclusively on *S*'s testimony and comes to know *p*, then *S* must also know *p*. Otherwise, the epistemic good required for transmission would be absent. Taken together, these cases underscore the central insight of the *CC* principles: inference and testimony do not create epistemic goods, such as knowledge or warrant, *ex nihilo*. Rather, such goods must already reside at the source in order for transmission to succeed.[6]

Given the flexibility and plausibility of the necessary conditions that *CC* principles establish for different epistemic goods across various contexts, we are now in a position to ask whether their initial appeal extends to newer and less familiar domains. Consider this other variant:

> *CCU* Necessarily, if (i) a receiver *R* holds certain propositional attitudes about a phenomenon solely on the basis of a sender *S*'s testimony, and (ii) *R* attains understanding of this phenomenon, then *S* understands the phenomenon in question.

At first glance, *CCU* seems to continue the pattern established by other *CC* principles, such as *CCK* and *CCT*. Like them, it excludes epistemic overdetermination and traces the transmission of epistemic goods back to their source. At the same time, however, *CCU* introduces a notable shift: it relocates the focus from knowledge to understanding, thereby forging a new and compelling link between testimony, understanding, and inquiry.[7]

This shift is particularly significant for two reasons. First, *CCU* captures a key intuitive requirement for the successful transmission of understanding. If *R* genuinely comes to understand a phenomenon solely through testimony, then it seems necessary that *S* already possesses that understanding. Otherwise, it would be mysterious how such an epistemic good could be passed on at all. In this way, *CCU* helps clarify the conditions under which understanding can be meaningfully conveyed.

Second, this shift also aligns with the widely held view that understanding, rather than knowledge, is the ultimate aim of inquiry (cf. Kvanvig 2003; Riggs 2003; Pritchard 2010; Baehr 2014). From this perspective, *CCU* not only builds

[5] Note that *CCT* (and related *CC* principles of testimony) must be distinguished from logically weaker "first-link" principles (cf. Faulkner 2000). According to the latter principles, necessarily, if (i) *R* believes that *p* solely on the basis of *S*'s testimony that *p*, (ii) *R* knows that *p*, and (iii) *R* and *S* are links in a testimonial chain with respect to *p*, then at least the first link in the chain knows that *p*. This principle requires that testimonial knowledge ultimately trace back to some knowledgeable source, but it does not require that every link in the chain itself possess knowledge. By contrast, *CCT* imposes a stronger constraint: it requires that if *R* knows *p* solely on the basis of *S*'s testimony, then *S* also know *p*. Hence, *CCT* assumes that knowledge can only ever be transmitted, never generated, through testimony. The weaker principle, however, allows for transmission to succeed despite intermediate links who lack knowledge, thereby leaving open the possibility of downstream acquisition of knowledge even when some upstream agents do not themselves possess it. In what follows, we focus on *CCT* precisely because it embodies the strict transmission model that the counterexamples in the next sections are designed to challenge. Thanks to an anonymous referee for requesting greater clarity on this point.

[6] For a discussion on memory and *CC* principles, see Luzzi (2019).

[7] By "inquiry" we refer to a structured, goal-directed process through which agents seek to acquire epistemic goods. This process involves not only individual cognitive effort but also engagement with the insights and testimony of others, particularly in complex or specialized domains.

on the insights of existing *CC* principles but also reinforces the idea that attaining one's epistemic goals often depends on the intellectual achievements of other epistemic agents. In other words, *CCU* brings into sharper focus our inherent *epistemic interdependence*: the recognition that epistemic agents rarely attain understanding in isolation but instead rely on a network of shared inquiry and testimonial exchange (cf. Hardwig 1985).

Having now outlined and extended the family of *CC* principles, we are ready to ask whether these principles are as plausible as they initially appear. To answer this, the next step will be to consider how *CC* principles of testimony apply to contemporary debates about AI. Before turning to that discussion, however, it is important to first examine *CCK*'s main shortcomings – especially its assumption, shared with *CCT* and *CCU*, that epistemic goods in these contexts are transmitted rather than generated.

3. Counter-closure…in trouble

Despite their intuitive appeal, *CC* principles face significant challenges. Take *CCK* for instance: several critics have convincingly argued that inferential knowledge can stem from premises that are *false* (cf. Warfield 2005; Klein 2007; Baggio 2025), *Gettierized* (cf. Luzzi 2010), or even *disbelieved* (cf. Murphy 2013). To better understand the impact of these challenges, we will now briefly examine a case where inferential knowledge arises from a false premise, illustrating how counterexamples to *CCK* can be readily constructed.

Consider the following scenario:

> *Fancy Watch*
>
> I have a 7 p.m. meeting and complete confidence in my expensive watch's accuracy. Having lost track of time and wanting to be punctual, I carefully check my watch and reason: "It is *exactly* 2:58 p.m.; therefore, I am not late for my 7 p.m. meeting." [...] I know my conclusion is true, but in fact, the actual time is 2:57 p.m., not 2:58 p.m. (Warfield 2005, 408)

This example highlights that, in certain circumstances, even when premise *p* is false, agent *S* can still competently infer and acquire knowledge of a conclusion *q*. The key insight here is that in many counterfactual scenarios where the same initial conditions of *Fancy watch* hold, if *S* later discovers that $\neg p$ after forming the belief in *q*, they could still competently rely on an alternative, epistemically close (proxy) premise to infer the same conclusion *q*.

For instance, if *S* can competently approximate time in the *Fancy watch* case, they could revise their initial belief to "It's approximately 2:58 p.m." upon learning that $\neg p$.[8] While this new belief is less precise, it is true and epistemically close enough to the original one to support the conclusion that *S* is not late. This suggests that inferential knowledge is more resilient than *CCK* assumes, challenging the idea that a false premise automatically undermines inferential knowledge.[9]

[8] In this case, *S* can rely on this approximate belief because every precise belief entails its approximation, and *S* is able to competently approximate the time on the basis of a more precise (though false) belief formed by looking at an accurate watch (cf. Baggio 2025 for an in-depth discussion).

[9] It is important to note that in cases like *Fancy Watch*, the formation of a new, approximately true belief depends heavily on first adopting the false belief. In other words, without accepting the false premise, *S* would struggle to arrive at the true approximation (cf. fn. 8). Still, one might object that *S* is already propositionally warranted – based on their initial evidence – in directly believing the approximate truth. On this view, it is this path, rather than the one based on the false belief, that does the real epistemic work. We submit, however, that this objection is unconvincing for two reasons. First, the objection overlooks a crucial asymmetry: why is that when an exact belief is true, *S* relies on it directly to draw further conclusions, but when the belief is false, *S* must instead rely on approximate truths? Since the only difference between this case and the original *Fancy watch* case lies in the accuracy of the watch itself, the objector must explain why such a

To further elaborate on this point, let's contrast the *Fancy watch* example with a classic Gettier-style case, where inferential knowledge breaks down.

> *Ford car*
>
> Nogot, Havit, and I are classmates. One day, I see Nogot driving a Ford, parking it in his garage, and claiming it's his car. Based on this, I conclude that Nogot owns a Ford, so someone in our class owns a Ford. Now, imagine that while it's true a classmate owns a Ford, it turns out not to be Nogot – it's Havit, about whom I have no beliefs regarding car ownership. (adapted from Lehrer 1965)

In this scenario, *S* believes "Someone in our class owns a Ford" based on the false premise that "Nogot owns a Ford". However, *S*'s inference does not qualify as knowledge here because *S* cannot competently derive a new true premise from the false belief. Why? Because in many counterfactual scenarios where the same initial conditions hold, *S* cannot competently revise their initial belief to infer the truth of the conclusion upon learning that $\neg p$. In effect, if *S* were to discover that "Nogot does not own a Ford", they would have no alternative warranted premise for believing "Someone in our class owns a Ford". In other words, their initial warrant cannot be transferred to a new, true (proxy) premise that could sustain their inferential knowledge once the original premise is compromised.

Therefore, the contrast between the *Fancy watch* and *Ford car* scenarios already highlights that not all instances of knowledge derived from ignorance are the same. In some cases, inferential knowledge is robust and survives despite a false premise; in others, it falls apart. This distinction highlights that *CCK*, which posits that false premises cannot generate inferential knowledge, is unconvincing. These differences reveal a significant flaw in *CCK*, rendering it an unpalatable principle for understanding the transmission of knowledge.[10]

In this section, we have briefly examined how knowledge can be generated from cases of ignorance. The next section will explore how this idea also applies to testimony cases.[11]

4. Breaking the chains

Let us consider the following scenarios:

> *Climate change denier*
>
> Lilith teaches science in a high school. The principal wants the students to learn about the causes and implications of climate change and asks Lilith to offer a seminar on the topic. Lilith consults the right sources […], understands the content, and gets ready for the seminar. On the basis of her readings, she asserts to her students that human activity and pollution are causally responsible for the rise in temperature on our planet, and she explains the details of the causal nexus between greenhouse gas emissions and global

worldly variation in the watch's setting should determine *S*'s method of inference (cf. Luzzi 2019). Second, as others have argued, this objection may be available – though only at significant cost – to the epistemic internalist about warrant, but not to the externalist (cf. Baggio 2025).

[10] There are some intermediate cases of knowledge from falsehoods that share relevant features with both cases described in this section (cf. Pritchard 2023). For the sake of brevity, we will not consider the discussion around these cases.

[11] The above considerations extend beyond *CCK*. Similar concerns arise in relation to other *CC* principles, particularly those governing epistemic warrant. For instance, inferential warrant does not always require that the premises themselves be warranted in order to confer warrant on a conclusion. A familiar case is *reductio reasoning*: even when the premises are neither warranted nor believed (or even disbelieved), the inference can still generate a warranted conclusion (cf. Murphy 2013, 2015; see also Baggio 2025).

> warming. The students judge Lilith to be an authority as far as science is concerned and do not hesitate to form corresponding beliefs on the basis of her testimony. Outside the classroom, Lilith is a climate change denier. She believes that there is no causal nexus between human activity and global warming, that global warming is a natural and unavoidable process, and that it will be naturally followed by a phase of temperature reduction. Despite her strong convictions, due mostly to her overall fatalism, she regards it as her duty as a teacher to share with her students the viewpoint of the majority of the members of the scientific community. (Malfatti 2019, 480)
>
> *AI Diagnostician*
>
> HealthBot-X is a cutting-edge, publicly accessible AI system designed to assist individuals in analyzing medical symptoms and conditions. Trained on extensive medical literature, patient records, and clinical studies, it can reliably detect patterns and correlations that human practitioners may miss. Notably, HealthBot-X itself lacks any medical beliefs. Instead, it relies purely on its learnings and ability to make predictions based on vast datasets, enabling it to uncover potential diagnoses that may have been overlooked. A compelling case highlights the system's potential: Tom's partner, desperate for answers, turned to HealthBot-X to help solve Tom's long-standing medical problem. Tom had been suffering from unexplained muscle weakness, chronic pain, and episodes of paralysis for over five years. Despite consulting various experts, no one could determine the cause of his condition. Due to concern, Tom's partner entered his symptoms, MRI data, and medical history into HealthBot-X. The AI quickly analyzed the data and suggested that Tom might have Stiff-Person Syndrome, a rare neurological disorder characterized by muscle stiffness and spasms, which can lead to debilitating disability if left untreated. HealthBot-X identified subtle patterns in the data that had been overlooked and recommended specific tests, including an anti-GAD antibody test, to confirm the diagnosis. When Tom's partner shared this diagnosis with his doctors, they carefully reviewed the AI's considerations and, based solely on these outputs, formed corresponding beliefs about Tom's condition, ultimately confirming the diagnosis. With proper treatment, Tom's symptoms significantly improved, offering him a better quality of life.[12]

As the attentive reader will have noticed, these cases closely parallel Lackey's (2008) well-known example of Stella, the creationist teacher. In that example, Lackey famously argues that we can acquire knowledge even from speakers who themselves do not possess it. For instance, although Stella does not believe in the theory of evolution, she can nevertheless foster knowledge of it in her students through testimony.

Now, all three cases, including Stella's, share a key structural feature: in each, a receiver *R* appears to gain testimonial epistemic goods from a sender *S* who lacks either knowledge or understanding of the relevant subject matter. This point should not be controversial. Stella rejects the theory of evolution; Lilith consciously sets aside her beliefs about climate change in order to maintain coherence with her broader worldview; and HealthBot-X, by its very nature, does not possess beliefs at all.

Building on this point, we can observe another important commonality: in these vignettes, *R* seems to gain testimonial epistemic goods not merely through transmission, where knowledge or understanding is passed along, but

[12] The interested reader can read a related news article here: Matteo Bassetti's Diagnosis with ChatGPT.

through testimonial acts that serve as *generative* epistemic sources.[13] As a matter of fact, if we can safely assume that *R* had never previously considered the proposition(s) conveyed by *S*, then these scenarios suggest that knowledge or understanding can be genuinely generated through testimony, even when *S* does not possess those epistemic goods themselves.[14]

To further highlight this point, consider the effects of swapping the protagonists across these scenarios. If HealthBot-X were placed in the role of Stella or Lilith, its epistemic role would remain unchanged. Although HealthBot-X lacks beliefs, it can competently process and convey information derived from reliable sources. In a classroom setting, it would still generate evolutionary knowledge or understanding, regardless of personal beliefs, much like Stella and Lilith. Conversely, if Stella or Lilith were substituted for HealthBot-X in the medical case, they could similarly generate

[13] For some, the case of Lilith may appear to be the least clear example of epistemic goods generation. Scholars like de Regt (2017), for instance, might be inclined to claim that Lilith possesses phenomenal understanding, and therefore that what occurs is a transmission – rather than a generation – of epistemic goods. This, of course, is a controversial point, and we do not aim to settle the debate here. Nevertheless, we believe Malfatti's (2019) account offers a promising perspective here. She argues that phenomenal understanding involves more than simply grasping (or entertaining) the content of, say, a theory, as appears to be the case with Lilith. Instead, it also requires internalizing that content and integrating it into one's own noetic profile. In effect, it seems quite counterintuitive to claim that Lilith *phenomenally* understands climate change while simultaneously denying it. A more plausible interpretation of her noetic profile is that she understands the *theory* of climate change, but not the *phenomenon* itself. According to Malfatti, this latter distinction is especially useful and we agree on this. Consider, for instance, the case of logic: a classical logician may understand the theory of paraconsistent logical validity, and a paraconsistent logician may likewise understand the theory of classical validity. However, we would not say that each possesses phenomenal understanding of logical validity in the same way. Rather, each phenomenally understands validity through the lens of their respective logical theory – classical or paraconsistent. This simultaneously explains why they both have good understanding of logical validity, but also why they disagree. Now, as a referee of this journal rightly notes, recent work in epistemology suggests that understanding may involve propositional attitudes beyond belief – such as acceptance or endorsement (cf. Elgin 2017; Malfatti 2022) – which in turn require stable dispositions to use content in making inferences. On this view, one might worry that, if acceptance alone suffices for phenomenal understanding, AI systems could plausibly be said to understand, since they can be programmed to perform such transitions. We submit, however, that phenomenal understanding – the kind relevant in Lilith's case – requires more than either belief or acceptance. What matters is not mere acceptance of content, but its integration into a wider noetic profile that is both internalized and available from a first-person perspective. By "internalized and available from a first-person perspective", we mean that the epistemic agent can reflect, endorse, and deploy the content as part of their own cognitive and doxastic landscape, rather than merely processing it as external information. AI systems may indeed display stable, functional transitions toward content that mimic acceptance. But absent a subjective standpoint and the richly embedded doxastic structure characteristic of human cognition such programmed dispositions fall short of the relevant epistemic good at stake. This conclusion, however, does not foreclose all avenues: it leaves open the more interesting question of whether other types of understanding might meaningfully be attributed to AI. Although this is an important issue, we will not pursue it here. For the sake of simplicity, the rest of the article will refer to beliefs as the relevant propositional attitude involved in phenomenal understanding. The careful reader, however, is free to replace this term with those discussed in this note. Thanks to an anonymous referee for requesting greater clarity on this delicate point.

[14] For the sake of clarity, it is important to address a potential objection at this stage. Discussions on testimony often distinguish between *reductionist* and *anti-reductionist* approaches (cf. Coady 1992). Simply put, reductionists treat testimony as a form of inference, where one premise typically concerns the speaker's trustworthiness or reliability and another states that the speaker has asserted that *p*. In contrast, non-reductionists consider testimony a non-inferential epistemic source, like perception. Given this distinction, one might argue that the critique of *CC* principles presented in the cases above applies only to non-reductionist accounts of testimony. This is because, for reductionists, testimony is never based solely on the utterance of *p*; it also depends on background beliefs, such as those previously mentioned. As a result, satisfying the no-overdetermination clause of *CCT* (or *CCU*) becomes problematic. One possible way to address this issue is to interpret the background beliefs required by reductionists as compatible with the no-overdetermination clause. In essence, this would involve distinguishing them from more identifiable beliefs that arise independently of the testimonial relationship in a given context (cf. Luzzi 2019). However, for those who find this move unconvincing, the problem may instead be seen as undermining the plausibility of reductionism. In effect, since *CC* principles of testimony appear *prima facie* compelling, the reductionist's inability to accommodate them weakens their position – unless, of course, they accept the line of argument developed here, which ultimately lends support to an anti-reductionist conception of testimony.

knowledge or understanding, provided they could engage with the medical data as HealthBot-X does – drawing inferences and citing reliable sources without necessarily holding personal beliefs in the diagnosis.

This latter observation naturally leads to another crucial consideration: these three cases are *epistemically isomorphic*. In other words, each epistemic element in one case has a structurally similar counterpart in the others. More precisely, this isomorphism suggests that the epistemic relations governing the generation and acquisition of epistemic goods in one testimonial exchange can be mapped onto another, preserving roles, dynamics, and epistemic effects. And crucially, this holds regardless of whether the subject is a human with personal disbelief, a human with a deviant noetic system, or an AI with no beliefs at all.

To make the isomorphism more vivid, we can break down the shared structural components. Each case features: (1) a sender *S* who conveys information, (2) a receiver *R* who forms certain propositional attitudes merely on the basis of that information, (3) the epistemic good generated – typically knowledge or understanding – and (4) the epistemic basis for the receiver's uptake (e.g., perceived reliability, authority, or inferential competence). What is striking is that across all three scenarios, these components interact in functionally equivalent ways, even as the internal doxastic states of the testifiers diverge radically.

Take the three examples in turn. In the Lilith case, the teacher lacks belief in anthropogenic climate change but is epistemically responsible in presenting the scientific consensus. Her students, treating her as an epistemic authority, acquire beliefs – and possibly understanding – through her testimony. In the AI case, HealthBot-X has no beliefs whatsoever but processes data with high inferential reliability; its outputs similarly lead to new understanding on the part of human users. Finally, in the Stella case (as per Lackey's original example), the creationist teacher imparts evolutionary theory with accuracy, despite her disbelief, and her students acquire scientifically grounded beliefs as a result.[15]

One might worry, however, that the AI case is less straightforward than the human analogues. Unlike Lilith and Stella, HealthBot-X does not operate in isolation from its designers and users. AI systems inevitably reflect the political and social contexts embedded in their design and training (cf. Flanagan et al. 2008; van de Poel and Kroes 2014; van de Poel 2020; see also Winner 1980). Put differently, one might suggest that beliefs, intentions, or forms of understanding are in some way always inscribed into the epistemic practices of systems like HealthBot-X. On this view, the system's functioning is not belief-free but rather bound up with the epistemic framings of its creators and users, who regard themselves as part of a shared epistemic community. We take this concern to highlight an important issue about the social and political embeddedness of AI systems. We agree that training inevitably inscribes epistemic practices, priorities, and framings into AI systems, much as human agents inherit patterns of reasoning and discourse from their communities. However, we draw a distinction between (i) *genuine beliefs*, which require deliberate endorsement and doxastic commitment, and (ii) *embedded patterns*, which may reflect training influences without any such endorsement. The

[15] There is an asymmetry between these cases that warrants closer scrutiny. While Lackey and Malfatti's cases involve testimony directed at laypeople, the AI case concerns testimony aimed at experts. While we believe this difference actually strengthens the AI case, one might argue that it undermines the comparison by highlighting a disanalogy between the scenarios. In response, we propose that Lackey and Malfatti's cases can be modified to make them more comparable to the AI case. Consider, for example, a figure like Will Hunting – a rebellious mathematical genius capable of absorbing and discovering virtually any body of knowledge simply by studying some books. His cognitive abilities seem to grant him an exceptional level of phenomenal understanding, even by the demanding standards of explanationism (cf. Kelp 2015; see also fn. 1). However, due to his rebellious nature, Will consistently refrains from committing to anything he reads or discovers. Despite this, he is tasked with presenting a lecture on his own discoveries, such as delivering a detailed proof of a long-unsolved mathematical conjecture. Although he refrains from committing to the truth of the assumptions or the logical inferential steps involved, he successfully conveys a correct proof to the experts in the audience through precise and compelling explanations. In such a case, it seems plausible to interpret each part of his presentation as an instance of *knowledge from ignorance*, and the lecture as a whole as a case of *understanding from non-understanding* involving testimony aimed at experts.

former is what we take to be philosophically relevant when asking whether an agent "believes" a proposition; the latter explains why AI outputs may mirror particular social and political contexts.

Moreover, recognizing this distinction allows us to situate AI alongside the human cases. Functional embedding in epistemic communities is not unique to AI. Human agents such as Stella and Lilith are likewise situated within social-epistemic frameworks, yet we do not count inherited but explicitly rejected claims as their beliefs. By parity of reasoning, HealthBot-X's social embedding shapes its functional competence but does not generate "beliefs" within the system itself. This, we suggest, aligns with our broader isomorphism claim: whether one "sets aside" beliefs consciously (Lilith) or operates without beliefs at all (HealthBot-X), they can still generate epistemic goods in testimonial contexts (like in inferential ones; see fn. 11).[16]

Hence, recognizing that this epistemic isomorphism holds carries two important implications. First, it reinforces the challenges to the assumption that *S*'s understanding or knowledge is necessary for competent testimony. This provides a unified challenge to traditional *CC* principles of testimony, such as *CCT* and *CCU*.

Second, the isomorphic nature of these cases suggests that *S*'s testimonial competence can be epistemically independent of their noetic system. That is, *S*'s ability to competently generate knowledge or understanding for *R* does not necessarily depend on the coherence, sincerity, or content of their own beliefs and understanding about the subject matter. What matters is their capacity to navigate, process, and convey information through effective epistemic practices.

As a matter of fact, in our examples, Stella, Lilith, and HealthBot-X are all able to deliver epistemically valuable testimony despite having no genuine belief in the content they convey. HealthBot-X, in particular, operates entirely outside the traditional human noetic framework: it lacks beliefs, intentions, or understanding in any phenomenally rich sense. Yet, under the right conditions, it reliably generates epistemic goods because of its ability to process data and follow reliable inference patterns drawn from its training on vast medical datasets, just as Stella and Lilith. This suggests that in some contexts, testimonial competence is less about an internal alignment between the sender's worldview and the content they convey and more about their ability to engage competently with established (scientific) practices.

To further clarify, in the above scenarios, Lilith sets aside her noetic commitments and employs the methods, sources, and inferential frameworks endorsed by the scientific community. Her testimonial competence stems from methodological alignment and disciplined engagement with reliable epistemic mediators, not from a sincere commitment to the truth of her statements. HealthBot-X, likewise, achieves testimonial competence for similar reasons. Its outputs are epistemically valuable because they result from reliable data processing and inference, not from belief or understanding in the human sense. Its competence arises from its role in a broader epistemic system.

Taken together, these reflections invite a refinement of our understanding of testimonial competence. Testimonial competence may be *role-relative* rather than *belief-relative*. What matters epistemically is not whether the source believes the testimony or integrates it into their worldview, but whether they effectively fulfill the role of an epistemic intermediary within a system of reliable practices (cf. Greco 2016).

However, this latter observation can be pushed even further. The isomorphic nature of these cases not only supports the claim that testimonial competence can be independent of personal belief or understanding; it also challenges the view that such agents are merely intermediaries – passive transmitters of knowledge or understanding that originated elsewhere. In effect, there is good reason to think that artifacts like AIs (and, by extension, even epistemic agents like Lilith and Stella), under certain conditions, can function as epistemic *originators*, genuinely initiating new chains of knowledge or understanding. In Tom's case, for instance, the AI did not simply relay existing knowledge or replicate

[16] Thanks to an anonymous referee for requesting greater clarity on this delicate point.

prior testimony. Instead, it synthesized information from disparate data points – symptoms, MRI scans, and medical history – and generated a diagnostic hypothesis that medical practitioners had forgotten or had not previously considered.

One might object, however, that this does not constitute a genuinely novel epistemic contribution, since the AI's reasoning still relies on patterns distilled from pre-trained human discoveries. Yet, other concrete cases offer a more compelling instantiation of epistemic origination in AI. For instance, Google DeepMind's AlphaFold unraveled protein structures that had long eluded biologists, effectively producing epistemic goods that were not easily derivable from existing human understanding at that time.[17] In such cases, the AI's output initiated a new epistemic trajectory.[18] In this sense, artifacts like HealthBot-X – and even more clearly, systems like AlphaFold – can function as epistemic originators rather than mere intermediaries (cf. Graham 2006).

To strengthen these conclusions, we now turn to potential defeaters. Specifically, we must examine whether there are compelling counterarguments that might dismiss these cases as epistemically irrelevant. Several objections warrant closer scrutiny.

4.1 Objection 1

One could argue that in all these cases, *S* merely acts as a conduit for information, not its true source. For example, with a human *S*, the real source of knowledge (or understanding) may be the teaching materials *S* relies on. Similarly, in the AI case, *R*'s understanding might stem not from the AI itself but from the anti-GAD test results its predictions generate.

At first glance, this objection could appear plausible. However, as Lackey (2008) argues, it is ultimately unconvincing for two independent reasons.

First, there is no compelling reason to assume that the author of the teaching materials must possess knowledge or an understanding of the subject matter. For instance, a creationist or climate change denier could have written the teaching materials, and if these agents lacked the relevant beliefs, they would also fail to transmit epistemic goods to *R* under either *CCT* or *CCU*. Yet, even proponents of Objection 1 would find it counterintuitive to claim that such materials cannot serve as effective sources of knowledge or understanding.

Second, Objection 1 overlooks an important counterfactual consideration: if *S* had the relevant beliefs, we would naturally credit them with helping *R* learn or understand the relevant facts. Since this counterfactual is true, dismissing *S*'s epistemic role based on their ignorance is question-begging and untenable. Therefore, *S* can play an indispensable role in generating knowledge (or understanding), which contradicts the claim that they are merely passive conduits.

[17] When we use the term "derivable" here, we mean it in both an epistemic and practical sense. The protein structures that AlphaFold predicted were theoretically derivable from existing scientific knowledge – the physical principles governing protein folding were well-established, and the experimental techniques existed. However, they were not practically derivable given the constraints of human cognitive resources, computational limitations, and the organizational structure of scientific research. The sources of this non-derivability are thus both epistemic and practical. On the one hand, the sheer complexity of the protein folding problem created what we might call an "epistemic bottleneck." While humans possessed the relevant background knowledge, the computational demands of applying this knowledge to predict complex structures from amino acid sequences exceeded human cognitive capacity. This represents a genuine epistemic limitation, which concerns not merely ignorance of principles, but also an inability to competently deploy existing knowledge (cf. Graham & Lyons 2021; Baggio 2025). On the other hand, the traditional scientific approach to determining protein structure required expensive and time-intensive experimental methods. What makes AlphaFold's contribution epistemically significant, then, is that it overcame both types of constraints simultaneously. It didn't simply transmit existing knowledge but synthesized vast amounts of biological data in ways that generated novel predictive capabilities – capabilities that were latent in the data but practically and epistemically unavailable to human researchers working within existing organizational and cognitive constraints. Thanks to an anonymous referee for requesting greater clarity on this delicate point.

[18] The interested reader can read a related news article here: Google's AI Lab, DeepMind, Offers 'Gift to Humanity' with Protein Structure Solution

Building on Lackey's considerations, we can see that they easily extend to the AI case. Consider, first, the claim that anti-GAD test data, rather than the AI itself, is the true source of testimony. One immediate problem with this view is that it overlooks a critical fact: raw data alone is epistemically inert. It must be interpreted in order to be meaningful.

Yet, even if we grant that the data has been interpreted, we can construct a parallel scenario to clarify the problem with Objection 1. For instance, imagine a lab technician who analyzes the data and writes a final report for doctors to assess. Suppose this technician deeply distrusts traditional medicine or machine-generated diagnoses, yet interprets the data reliably. According to *CCT* and *CCU*, they, too, would fail to transmit epistemic goods. However, denying their epistemic role in this scenario feels equally counterintuitive.

Additionally, one should note that the AI's output was instrumental in prompting the doctors to consider using the anti-GAD test – something they would likely not have done on their own (cf. fn. 17). This indicates that the AI played a crucial epistemic role in shaping the doctors' understanding of Tom's condition. Their eventual recognition of the test's relevance and reliability did not arise from an independent investigation. Instead, it emerged from a chain of reasoning initiated by the AI's output. Notably, their understanding of the diagnostic value of the test originates from a process that begins with a source – the AI – that lacks phenomenal understanding. In this sense, their resulting insight qualifies as a genuine form of testimonial understanding emerging from a non-understanding source. Hence, this challenges the assumption that testimony must always be grounded in phenomenal understanding to be effectively conveyed.

Further, it is also important to recognize here that understanding is a gradable phenomenon. If doctors derive a new level of understanding (understanding$_2$) from the test's results at time t_1, this does not conflict with our central thesis: that doctors can initially develop a basic level of understanding (understanding$_1$) solely from the AI at t_0. What matters is that understanding$_1$, generated by the AI, forms the foundation for understanding$_2$, which the doctors can later achieve. Crucially, understanding$_1$ is not simply transmitted to the doctors – it is produced by the AI itself, as the case presented here suggests.

Finally, the same counterfactual considerations Lackey presents in the human *S*'s context also apply here. After all, if *S* were a human being (like Stella or Lilith) possessing the requisite propositional attitudes, we would naturally credit them with helping *R* understand the facts about Tom's condition. Since this counterfactual also holds, it reinforces the argument that AI plays an essential epistemic role in generating understanding. Therefore, Objection 1 ultimately fails – even in the AI case.

4.2 Objection 2

One could argue that, in all these cases, *R* does not acquire the relevant epistemic goods through *S*'s testimony. Specifically, this objection challenges whether clause (ii) of *CCT* and *CCU* is satisfied, thereby questioning the applicability of these principles to the cases at hand.

However, Objection 2 ultimately proves unconvincing (cf. Lackey 2008). The crucial point in response is that *S*, in each case, draws on highly reliable sources. To see this more clearly, consider a scenario in which a non-deviant *S*, relying on the same reliable data, conveys the relevant information to *R* in the same manner. In such a scenario, it seems highly plausible that we would attribute knowledge or understanding to *R*. By analogy, the same reasoning extends to the AI-related cases under discussion: just as *R* would acquire epistemic goods in the non-deviant scenario, so too it is reasonable to say that *R* acquires them in the AI case. Thus, Objection 2 ultimately falls short – even in the AI case.

4.3 Objection 3

Finally, one could argue that, in all these cases, *S* is not testifying at all. This objection reframes the discussion by questioning whether condition (i) of *CCT* and *CCU* is satisfied – namely, whether *S* is genuinely testifying in the first place.

Unlike the previous objections, this one presents a more nuanced challenge that requires deeper consideration. The difficulty arises because *S* must meet at least three seemingly necessary conditions to testify that *p* genuinely:

1. The testifier must *intend* to provide testimony (cf. Pritchard 2004; Fricker 2015).
2. The testifier must be *subject to normative* assessment (cf. Goldberg 2012; Butlin & Viebahn 2025).
3. The testifier must participate in *trust* relations (cf. Faulkner 2011).

Taken together, these criteria suggest that while Stella and Lilith plausibly qualify as genuine testifiers, HealthBot-X falls short. Yet this conclusion cannot be drawn too hastily. Before accepting the objection, it is important to examine more carefully how these three constraints operate and whether they truly exclude HealthBot-X from the category of testifiers.

To begin with, it is essential to recognize that the strongest argument against AI-based testimony stems from condition 2 of the above list. In effect, while testimony is traditionally tied to normative accountability, several scholars now argue that it does not necessarily require *intentionality* (cf. Coady 1992; Lackey 2008) or *trust relations* (cf. Graham 2012; Simion & Kelp 2020). If so, the exclusion of AI-based testimony on the basis that AIs lack these features remains open to debate. The focus, therefore, should be on whether normative accountability – broadly construed – can apply to AI systems.

Moreover, it is crucial to note that the necessary conditions outlined in the list above are highly *anthropocentric* (cf. Freiman 2024a, 2024b; He & Yang *forthcoming*). Indeed, all these conditions presuppose that only human agents can engage in testifying, since only humans are, in principle, recognized as capable testifiers. This raises a further concern: if testimony is conceived as automatically excluding non-human agents like AIs, Objection 3 risks begging the question. This raises a more fundamental issue: can we broaden the concept of testimony to include non-human agents without undermining its core features?

Finally, and closely related to the preceding points, a promising way to address the challenges discussed above is to distinguish between *prescriptive* and *non-prescriptive* norms. In brief, prescriptive norms guide behavior by being represented or internalized by an epistemic agent. Non-prescriptive norms, by contrast, regulate a system's traits based on performance rather than internalized standards. For example, drawing on Burge (2003), Graham (2013) argues that norms can also be understood as evaluative, non-prescriptive benchmarks.

To see how this distinction works in practice, consider the case of a biological trait, such as the heart (cf. Millikan 1984). Its primary function is to pump blood reliably. According to a widely accepted view, the heart acquired this function because its blood-pumping ability was beneficial to organisms that possessed it in the past, which led to its preservation through natural selection.

Given this background, let us now ask ourselves: can a heart be held responsible for its failings? At first glance, the answer seems obvious. A heart does not internalize or represent norms, and thus cannot be held accountable in the traditional, prescriptive sense. Yet, there remains an important respect in which hearts can still be normatively evaluated. To see this, consider the following questions:

I. Did the heart pump blood? – *Did it fulfill its function*?

II. Did the heart pump blood normally? – *Did it operate normally according to nature's design*?

III. Did the heart pump blood reliably? – *Did it consistently fulfill its function in normal conditions*?

Before pursuing these questions further, some clarifications are in order. First, the use of terms like "function" and "normal" follows the technical philosophical framework developed by Burge (2003) and Graham (2011, 2019), which differs significantly from everyday usage. Within this framework, a "function" refers to an effect of a trait that explains why the trait exists – typically because earlier instances were selected for producing that effect, whether through evolutionary history or learning processes. "Normal functioning", by contrast, refers to operating in the manner that past instances did when they produced the effect responsible for their selection, while "normal conditions" are the circumstances under which that selection originally took place (cf. Graham 2011, 11). Second, this account is inherently pluralistic and context-sensitive, since functions and normal conditions are determined by the distinctive history of each system rather than by universal standards.

With these clarifications in mind, we can now return to the main point. According to Graham (2013), if a heart satisfies any of the evaluative criteria listed in the questions above, it can be considered a good example of its kind. In other words, biological traits can be normatively assessed based on their performance, even in the absence of intentionality or internalized norms. Their performance can still be meaningfully evaluated using non-prescriptive standards of success.

This insight, then, presents a promising opportunity for reconceptualizing the normative basis of testimony in a way that includes AI systems. Suppose we can accept that non-prescriptive norms can serve as a legitimate form of normative assessment. In that case, it becomes possible to regard AI outputs as falling within a broader category of testimony-like acts. Similarly to how we assess whether a heart is functioning well by examining whether it pumps blood effectively and reliably according to nature's design, we can similarly evaluate whether an AI system like HealthBot-X is "testifying well" by asking whether it outputs health-related information accurately, in accordance with its specifications, and consistently under normal conditions (cf. van de Poel 2020).

Crucially, this approach also enables us to sidestep the anthropocentric assumptions inherent in traditional conceptions of testimony, which typically require features such as prescriptive normative assessment – features that AI currently lacks. By appealing to non-prescriptive norms, we are no longer forced to reject AI-based contributions outright on the basis that AIs do not "intend" to testify or cannot be held responsible in the moral or legal sense. Instead, the focus shifts to the quality and reliability of their outputs, which aligns more closely with the broad epistemic function of testimony: to provide information that others can rely on in forming beliefs.

Moreover, adopting this framework avoids the drawback of begging the question against AI testimony. It does not assume from the outset that only humans can testify, but instead reexamines what it means to testify from the ground up. If what matters most is that a source provides information in a normatively assessable way – whether through internalized norms or functional benchmarks – then AIs can, at least in principle, meet the minimal conditions required for normative evaluation. This suggests a more inclusive and functionally grounded account of testimony, one that preserves its epistemic purpose without arbitrarily excluding epistemic artifacts.

Finally, it is also worth emphasizing that epistemic agency, as we understand it here, should not be confused with the mere appearance of agency. What matters for testimonial competence is not whether a system can convincingly imitate human-like responses or pass tests of apparent rationality (cf. Pantsar 2025), but its ability to participate effectively in epistemic practices through outputs derived from processes that reliably track truth *via* appropriate learning mechanisms. This sets apart genuine epistemic sources from systems that may display complex behaviors through simple rules or reactions but lack the necessary connection between their outputs and truth-tracking processes, such as ant colonies solving maze-like problems with pheromone trails or slime molds seemingly navigating obstacles through biochemical processes.[19]

That said, it is sufficient for the purposes of this article to recognize that, like biological traits, AI systems come with specific (design) functions – functions they fulfill to varying degrees of success. While the mechanisms through which AIs acquire these functions may differ from evolutionary processes, their role as epistemic artifacts with function-based norms is well-established (cf. Burge 2010; Graham 2019). Given this, we can conclude that excluding AI from the category of potential testifiers lacks justification. Objection 3, therefore, does not hold – even in the case of AI.

5. Conclusions

The epistemic parallels discussed in this paper, together with our extended analysis of the three objections, indicate that the cases of the *Creationist Teacher*, *Climate Change Denier*, and *AI Diagnostician* pose serious challenges to *CC* principles like *CCT* and *CCU*. This conclusion carries two significant implications. First, it strengthens and extends existing critiques of the widely accepted view that the intuitive assumptions underlying these *CC* principles are necessary for epistemic sources like testimony and inference. Second, it supports the idea that if we accept that human agents can acquire knowledge or understanding from sources that are ill-informed yet still competent, then we should also be open to the possibility that artificial systems – such as AIs – can produce genuine epistemic goods through testimony, even in ways that go beyond current expectations.

Conflict of Interest: The author declares no conflict of interest.

Acknowledgments: I am sincerely grateful to Federica Malfatti for the many enlightening conversations we have shared on these topics. I am also indebted to the two anonymous referees of this journal for their insightful comments and constructive suggestions. Finally, I wish to thank the journal's Editor-in-Chief, Eric Kerr, for his guidance and support.

Funding: Funded by the European Union – Next Generation EU, Mission 4, Component 1, CUP I53D23006890001.

References

Audi, R. (1997). The Place of Testimony in the Fabric of Knowledge and Justification, *American Philosophical Quarterly*, 34, 405-422.

Audi, R. (2010). *Epistemology: A Contemporary Introduction to the Theory of Knowledge* (3rd edition). Routledge.

Baehr, J. (2014). Sophia: Theoretical Wisdom and Contemporary Epistemology. In K. Timpe & C. Boyd (Ed.), *Virtues and their Vices* (1st ed., p. 303–323). Oxford University Press.

Baggio, M. (2025). Knowledge from Falsehoods Reconsidered. *Episteme*, 1-26.

Bender, E. M., Gebru, T., McMillan-Major A., Shmitchell S. (2021). On the dangers of stochastic parrots: can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610-623.

Bengson, J. (2015). A Noetic Theory of Understanding and Intuition as Sense-Maker. *Inquiry: An Interdisciplinary Journal of Philosophy*, 58(7-8), 633-668.

Burge, T. (1993). Content preservation. *Philosophical Review*, 102(4), 457-488.

Burge, T. (2003). Perceptual entitlement. *Philosophy and Phenomenological Research*, 67(3), 503-48.

Burge, T. (2010). *Origins of Objectivity*. Oxford University Press.

[19] A referee raises the important question of what distinguishes human and AI epistemic agency if the cases are truly isomorphic. While the epistemic isomorphism we have discussed in this paper holds in testimonial contexts, this does not collapse all distinctions between human and AI agency. The key criterion remains whether competence emerges from learning processes that reliably track truth over time, rather than from mere functional similarity or behavioral mimicry.

Butlin, P., & Viebahn, E. (2025). AI assertion. *Ergo*, 12, 969–988.

Coady, C. A. J. (1992). *Testimony: A Philosophical Study*. Oxford University Press.

de Regt, H. W. (2017). *Understanding Scientific Understanding*. Oxford University Press.

Elgin, C. Z. (2017). *True Enough*. MIT Press.

Faulkner, P. (2000). The Social Character of Testimonial Knowledge. *The Journal of Philosophy*, 97, 581-601.

Faulkner, P. (2011). *Knowledge on Trust*. Oxford University Press.

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying values in technology. Theory and practise. In J. Van den Hoven & J. Weckert (Eds.), *Information technology and moral philosophy* (1st ed., pp. 322–353). Cambridge University Press.

Floridi, L. (2023). AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, 36, 15.

Freiman O. (2024a). Analysis of Beliefs Acquired from a Conversational AI: Instruments-based Beliefs, Testimony-based Beliefs, and Technology-based Beliefs. *Episteme*. 21(3), 1031-1047.

Freiman, O. (2024b). AI-Testimony, Conversational AIs and Our Anthropocentric Theory of Testimony. *Social Epistemology*, 38(4), 476-490.

Fricker, E. (2015). How to Make Invidious Distinctions Amongst Reliable Testifiers. *Episteme*, 12(2), 173-202.

Goldberg, S. C. (2012). Epistemic Extendedness, Testimony, and the Epistemology of Instrument-based Belief. Philosophical Explorations, 15(2), 181-97.

Graham, P. J. (2011). Epistemic Entitlement. *Noûs*, 46(3), 449-482.

Graham, P. J. (2012). Testimony, Trust, and Social Norms. *Abstracta*, 6(S6), 92-116.

Graham, P. J. (2013). Warrant, Functions, History. In A. Fairweather & O. Flanagan (Eds.), *Naturalizing Epistemic Virtue* (1st ed., p. 15-35). Cambridge University Press.

Graham, P. J. (2019). Why is Warrant Normative? *Philosophical Issues*, 29(1), 110-128.

Graham, P. & Lyons, J. (2021). The Structure of Defeat: Pollock's Evidentialism, Lackey's Framework, and Prospect for Reliabilism. In J. Brown and M. Simion (Eds.), *Reasons, Justification, and Defeat* (1st ed., pp. 39-68). Oxford University Press.

Greco, J. (2014). Episteme: Knowledge and Understanding. In K. Timpe and C. A. Boyd (Eds.), *Virtues and their Vices* (1st ed., p. 285–302). Oxford University Press.

Greco, J. (2016). What is transmission? *Episteme*, 13(4),481-498.

Grimm, S. (2014). Understanding as Knowledge of Causes. In A. Fairweather (Ed.), *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science* (1st ed., p. 329-345). Springer

Grodzinsky, F., Miller, K. W., & Wolf, M.J. (2020). Trust in artificial agents. In J. Simon (Ed.), *The Routledge handbook of trust and philosophy* (1st ed., p. 298-313). Routledge.

Hardwig, J. (1985). Epistemic dependence. *Journal of Philosophy*, 82(7), 335-349.

He, J., Yang, C. (*forthcoming*) Testimony by LLMs. *AI & Society*.

Kelp, C. (2015). Understanding phenomena. *Synthese*, 192, 3799-3816.

Khalifa, K. 2017, *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press.

Klein, P. (2008). Useful false beliefs. In Q. Smith (Ed.), *Epistemology: new essays* (1st ed., p. 25-63). Oxford University Press.

Kvanvig, J. L., (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.

Lackey, J. (2008). *Learning from Words*. Oxford University Press.

Lehrer, K. (1965). Knowledge, truth and evidence. *Analysis*, 25, 168-175.

Lipton, P. (2009). Causation and Explanation. In H. Beebee, C. Hitchcock & P. Menzies, *The Oxford Handbook of Causation*. Oxford University Press.

Luzzi, F. (2010). Counter-Closure. *Australasian Journal of Philosophy*, 88(4), 673-683.

Luzzi, F. (2019). *Knowledge from Non-Knowledge*. Cambridge University Press.

Malfatti, F. I. (2019). Can Testimony Generate Understanding? *Social Epistemology*, 33(6), 477-490.

Malfatti, F. I. (2022). Understanding phenomena: From social to collective? *Philosophical Issues*, 32(1), 253-267.

Malfatti, F. I. (2025). ChatGPT, Education, and Understanding. *Social Epistemology*, 1-15.

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

Murphy, P. (2013). Another Blow to Knowledge from Knowledge. *Logos and Episteme,* 4(3), 311-317.

Murphy, P. (2015). Justified belief from Unjustified belief. *Pacific Philosophical Quarterly*, 98, 602–617.

Pantsar, M. (2025). Intelligence is not deception: from the Turing test to community-based ascriptions. *AI and Society*, 40(5), 4065-4077.

Pritchard, D. (2004). The epistemology of testimony. *Philosophical Issues*, 14(1), 326-348.

Pritchard, D. (2010). Knowledge and Understanding. In D. Pritchard, A. Millar, & A. Haddock (Eds.), *The Nature and Value of Knowledge: Three Investigations* (1st ed., p. 1-88). Oxford University Press.

Pritchard, D. (2023). Knowledge from Error and Anti-Luck Virtue Epistemology. In R. Borges & I. Schnee (Eds.), *Illuminating Errors* (1st ed., p. 93-103).

Riggs, W. (2003). Understanding 'Virtue' and the Virtue of Understanding. In M. DePaul & L, Zagzebski (Eds.), *Intellectual Virtue: Perspectives From Ethics and Epistemology* (1st ed., p. 203-226). Oxford University Press.

Schechter, J. (2017). No Need for Excuses. Against Knowledge-First Epistemology and the Knowledge Norm of Assertion. In J. A. Carter, E. C. Gordon & B. W. Jarvis (Eds), *Knowledge First. Approaches in Epistemology and Mind* (1st ed., p. 132-161). Oxford University Press.

Simion, M. & Kelp, C. (2020). How to be an anti-reductionist. *Synthese*, 197(7), 2849-2866.

van de Poel, I., & Kroes, P. (2014). Can technology embody values? In P. Kroes & P.P. Verbeek (Eds.), *The moral status of technical artifacts* (1st ed., pp. 103–124). Springer.

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. *Minds & Machines* 30, 385–409.

Warfield, T. (2005). Knowledge from falsehood. *Philosophical Perspectives,*19(1), 405-416.

Williamson, T. (1994). *Vagueness*. Routledge.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press.

Williamson, T. (2007). *The Philosophy of Philosophy*. Blackwell.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109, 121–136.