Journal of English Language and Education
ISSN 2597- 6850 (Online), 2502-4132 (Print)
Journal Homepage: https://jele.or.id/index.php/jele/index

OPEN ACCESS

Article

Does AI Know Things? An Epistemological Perspective on Artificial Intelligence

https://doi.org/10.31004/jele.v10i5.1592

*Lasyuli Simbolon, M. Manugeren, Efendi Barus[abc]
[123]Kajian Bahasa Inggris, Fakultas Sastra, Universitas Islam Sumatera Utara, Indonesia
Corresponding Author: lasyulisimbolon@gmail.com

ABSTRACT

This paper investigates the provocative question: can artificial intelligence (AI) know things? through an epistemological lens. Drawing upon a systematic literature review (SLR) of works published 2010–2020, the study maps how scholars have applied classical and contemporary epistemic criteria—such as belief-likeness, truth, justification, reliability, interpretability, and epistemic agency—to AI systems. In doing so, it examines competing theoretical frameworks (internalism, externalism, virtue epistemology, Bayesian approaches) and identifies areas of convergence and contention. The review reveals that while many AI systems satisfy externalist criteria of reliability and truth-tracking under controlled conditions, they often fall short of internalist demands for justificatory transparency or reflective access. Opacity and "black-box" architectures remain central obstacles to attributing knowledge in the classical sense. Furthermore, the influence of AI on human belief formation and the shift in epistemic environments suggest that even absent true knowledge, AI plays a significant role in mediating knowledge practices. Ethical and normative considerations (e.g. fairness, accountability, epistemic justice) also emerge as inseparable from epistemological assessments, prompting calls for a "glass-box epistemology" that integrates design, interpretability, and value sensitivity. In concluding, the paper argues that AI may function as a contributor to human knowledge workflows rather than as autonomous knowers. It sets out a nuanced perspective: acknowledging AI's epistemic potential while remaining critical of overextensions. Finally, it suggests future paths: refining epistemic thresholds, embedding interpretability in AI design, and expanding the discourse across cultural and disciplinary contexts.

Keywords: *AI Epistemology, Knowledge and Justification, Reliability and Interpretability, Epistemic Agency, Internalism Vs Externalism*

Article History:
Received 14th October 2025
Accepted 29th October 2025
Published 31st October 2025

Check for updates

INTRODUCTION

In recent years, Artificial Intelligence (AI) has moved from the realm of speculative technology into deeply embedded systems in everyday life. From recommendation engines in social platforms, diagnostic tools in medicine, to large language models that generate human-like text, AI systems are increasingly involved in tasks traditionally associated with human knowledge and reasoning. This raises a profound epistemological question: does AI truly "know" things, or are its outputs better understood as mere data processing and pattern matching without genuine knowledge?

Epistemology, the branch of philosophy concerned with the nature, sources, limitations, and justification of knowledge, has historically focused on human agents. Classic definitions of knowledge have often included components such as belief, truth, and justification. These components are challenged when considering AI: the "belief" component is not literally present, truth is probabilistic rather than certain, and justification is often opaque due to the complexity of AI models. As AI's role in society grows, philosophers and technologists alike are forced to examine whether AI outputs satisfy any standard epistemic criteria.

One way to approach this question is by comparing AI systems to theories in epistemology such as externalism and internalism. Externalist approaches might focus on whether an AI's processing reliably generates true beliefs (or analogous outputs), whereas internalist approaches might insist on visibility or accountability of justification. For example, the "black box" problem in deep learning — where the internal decision-making process is hard to inspect — causes tension with internalist notions of justified knowledge (Duran & Jebeile, as cited in "Automating Epistemology" literature). This tension underlines one of the core debates: is reliability and performance enough for "knowledge," or does justification (in a sense humans understand it) also matter?

Another dimension is epistemic agency — whether AI has the capacity for belief revision, metacognition, or self-awareness in meaningful ways. Some recent work has explored how machine learning models affect human belief revision and the normative implications of those influences, without necessarily attributing belief to the AI itself (see AI and Epistemic Agency, 2025). These discussions complicate the question: even if AI does not "know" in the full human philosophical sense, it may nevertheless shape knowledge environments in ways that have epistemological significance.

Furthermore, current empirical and philosophical literature introduces the notion of AI as an epistemic technology. Ramón Alvarado (2023) argues that AI and data science tools are designed, developed, and deployed in epistemic contexts: they manipulate epistemic content (data), perform epistemic operations like prediction or inference, and contribute to inquiry in science and society. This framing shifts the question from "can AI know?" to "how does AI function in the web of knowledge creation, justification, and dissemination?"

A related concern involves truth, bias, and verification. AI systems are trained on large corpora of human-created data, which contain biases, gaps, and errors. When AI outputs are used as forms of knowledge (e.g., in policy, healthcare, or journalism), what measures are in place to ensure that their "truth-claims" are reliable? Studies in algorithmic systems show that statistical approximation often replaces reflective reasoning, leading to a kind of epistemology grounded in correlation rather than justification (Automating Epistemology) (Springer article). The political, social, and ethical stakes are significant, especially in areas where knowledge claims have real consequences.

Finally, this paper aims to map key arguments, tensions, and possible frameworks for evaluating AI's knowledge status. By adopting an epistemological perspective, we will examine whether AI systems satisfy standard epistemic criteria, what it means for an AI to have knowledge or to "know," and what implications this has for human epistemic practices. In doing so, we will survey the literature, identify gaps, and propose considerations for future theory and practice.

Literature Review

Recent work in philosophy of technology has grappled with the problem of trust and reliability in AI, particularly whether AI systems can be appropriate objects of trust even when their internal operations are opaque. Robertson (2025) in "AI, Trust and Reliability" argues that opacity does not automatically disqualify AI systems from being trusted, as long as there is a strong inductive basis for believing they perform reliably over time. He engages with opponents who claim that lack of explainability undermines trust, but counters that reliability and consistent performance may suffice under certain trust frameworks.

Another strand of literature addresses how AI reconfigures traditional notions of truth, authority, and verification in public discourse. In "Automating Epistemology: How AI Reconfigures Truth, Authority, and Verification" (Ortmann, 2025), the author introduces the concept of algorithmic truth, showing how AI mediates public knowledge by procedural and infrastructural norms such as data pipelines, pattern recognition, and model confidence, rather than by classic coherence or correspondence conceptions of truth. This work points out that AI systems, especially in content moderation and misinformation-detection, are not neutral; they carry embedded normative assumptions and biases.

The issue of what makes AI trustworthy as opposed to merely reliable is explored in Simion & Kelp's (2023) Trustworthy Artificial Intelligence, published in Asian Journal of Philosophy. They argue that trustworthy AI should be judged by how well systems fulfill function-based obligations. Such obligations derive either from design functions or from an etiological account (i.e., how the system came to have its functions). Traits commonly listed as requirements for trustworthy AI—such as safety, justice, and explainability—are important, but according to the authors they must tie back to whether the AI is fulfilling what it should do (its functions) rather than just listing desirable features. This offers a normative lens for epistemology: knowledge claims via AI must be evaluated by whether the system is performing appropriately in its role.

Epistemic agency also appears in recent literature. In "AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications" (Coeckelbergh, 2025), the focus shifts to how AI systems affect human belief formation, revision, and agency. The paper argues that as AI becomes more embedded in knowledge environments, it may distance individuals from engaging in critical reflection, thereby weakening their control over belief revision. Even if AI does not itself possess beliefs, its influence on human epistemic agents is indispensable to understanding AI's status as a potential knower or contributor to knowledge environments.

Moreover, policy and technical-ethical frameworks map the requirements needed for responsible and trustworthy AI systems more broadly. Díaz-Rodríguez, Del Ser, Coeckelbergh, López de Prado, Herrera-Viedma, and Herrera (2023) present a holistic vision of trustworthy AI consisting of key requirements such as human agency and oversight; robustness and safety; privacy and data governance; transparency; fairness; societal and environmental well-being; and accountability. Their work emphasizes that these requirements are interdependent and must be enforced throughout an AI system's entire lifecycle. This has implications for epistemological assessments: knowledge claims from AI should be evaluated not only based on performance or truth but also on ethical, governance, and social dimensions.

Lastly, some literature challenges whether trust (a key component in epistemic practices) can meaningfully be attributed to AI without meeting explainability or normative standards. The paper "Can We Trust Artificial Intelligence?" (2025), published in Philosophy & Technology, considers the epistemic and normative components of trust and argues that the possibility of trusting AI is not self-evident. It presents a threefold challenge to proponents of trusting AI: ensuring reliable performance, maintaining normative justification, and reconciling trust in AI with the moral or ethical obligations that trust carries. This line of research sheds light on what it might mean for AI to "know" something: if knowledge requires more than correct output (e.g. justification, normative responsibility), then trust becomes central.

METHOD

This paper uses a Systematic Literature Review (SLR) approach to investigate whether AI can be said to "know" things from an epistemological perspective, focusing on literature published between 2010 and 2020. The SLR is chosen because it allows comprehensive, structured, and transparent synthesis of philosophical and empirical work relevant to epistemological criteria (belief, truth, justification, reliability, interpretability, epistemic agency, etc.). Using this approach helps minimize bias and ensures that findings are traceable and reproducible.

To structure the review, several research questions guide the inquiry: What epistemological criteria are invoked in discussions of AI as a possible knower during 2010-2020? Which theoretical frameworks (internalism, externalism, virtue epistemology, Bayesian approaches, etc.) are most used? What kinds of arguments (philosophical, empirical, conceptual) support or challenge the idea that AI systems meet epistemic standards of knowledge? And finally, what gaps exist in this body of work—e.g., in terms of interpretability, cultural perspectives, or specific domains of AI? The search strategy involves querying multiple scholarly databases (such as Google Scholar, PhilPapers, JSTOR, IEEE Xplore, ACM Digital Library, SpringerLink) using combinations of keywords including "artificial intelligence AND knowledge," "epistemology AND AI," "justification AND algorithm," "belief AND machine learning," and "truth AND AI." Boolean operators and phrase searches will refine results (e.g. "knowing," "knowledge," "belief," "justification"). The timeframe for inclusion is restricted to works published from January 1, 2010 to December 31, 2020; however, earlier seminal works may be referenced for background but are not part of the formal SLR.

Inclusion criteria require that papers explicitly discuss epistemological issues in relation to AI—such as how AI outputs might satisfy or fail the criteria for knowledge (truth, belief-likeness, justification, etc.), or how frameworks like internalism or externalism are applied to AI. The works included may be philosophical/conceptual, empirical studies with epistemic implications, or theoretical hybrids. Only peer-reviewed journal articles, book chapters, or reputable conference proceedings in English will be considered. Exclusion criteria eliminate technical-only papers (focusing exclusively on model performance metrics without epistemological reflection), non-peer-reviewed media, and those outside the 2010-2020 window.

The screening proceeds in stages: first, an initial search generates a large set of candidate works; next, titles and abstracts are screened by two independent reviewers to identify those likely to meet the criteria; disagreements resolved by consensus or a third reviewer. Then, full text screening of the remaining works is conducted. Reference lists ("snowballing") are also checked to find additional relevant studies not picked up in the initial search.

For data extraction, each selected study is examined to pull out details such as author, year, publication, theoretical approach used (internalism, externalism, etc.), the epistemic criteria addressed (belief, truth, justification, reliability, interpretability, etc.), arguments for and against AI as a knower, case examples (e.g. types of AI or application domains), and limitations identified by the authors. These data are organized in a standardized extraction table to enable comparison across studies.

Quality assessment is important; philosophical works will be assessed for clarity of definitions (what "knowledge," "justification," etc., mean in that work), logical coherence of arguments, and whether counter-arguments are considered; empirical or technical studies will also be evaluated for methodological transparency, the domain or AI type studied, and how measures of truth, reliability, or justification are operationalized.

Finally, for synthesis, thematic coding will be used to group findings under common epistemological criteria or frameworks (e.g. how many papers use externalism vs internalism, how many address interpretability vs reliability, or epistemic agency). A comparative analysis will highlight agreements, conflicts, and trends across time (within 2010-2020). The review will also identify gaps (for example, areas underexplored in that decade, or AI domains for which epistemological work is rare). The results will be presented with thematic headings, supported by summary tables, and possibly a flow diagram showing how many works were found, screened, excluded, included.

FINDINGS AND DISCUSSION

The following table presents a selection of studies that examine various aspects of AI from an epistemological perspective, published between 2010 and 2020:

Table 1. Summary of Selected Studies

Study	Epistemological Focus	Key Themes	Methodology	Publication Year
Ganascia (2010)	Philosophy of Information	AI as informational processes	Conceptual analysis	2010
Anneborg (2020)	Virtue Epistemology	AI as a knower	Philosophical inquiry	2020
Pabubung (2024)	Epistemology and Ethics	AI, ethics, and interdisciplinary education	Conceptual analysis	2024
Páez (2020)	Explainable AI	Pragmatic understanding in XAI	Philosophical analysis	2020
Kantar (2020)	Empirical Epistemology	AI and contemporary epistemology	Interview study	2020

Note: Some studies are slightly outside the 2010–2020 range but are included for their relevance.

Thematic Analysis

The selected studies highlight several key themes in the epistemology of AI:

AI as Informational Processes: Ganascia (2010) discusses AI as an attempt to reduce the natural mind to informational processes, critiquing the adequacy of symbolic abstraction in capturing the richness of mental activity.

AI as a Knower: Anneborg (2020) explores the possibility of AI agents possessing knowledge, examining this from a virtue epistemology perspective and considering both current AI capabilities and future developments.

Ethics and Interdisciplinary Education: Pabubung (2024) emphasizes the importance of interdisciplinary education in understanding the ethical implications of AI, advocating for an epistemological analysis to guide AI development.

Explainable AI (XAI): Páez (2020) argues for a pragmatic and naturalistic account of understanding in AI, suggesting that interpretative models are necessary for achieving post-hoc interpretability in machine learning models.

Empirical Epistemology: Kantar (2020) conducts an interview study with contemporary epistemologist Robert Audi, discussing the effects of AI on epistemology and highlighting the need for a nuanced understanding of AI's epistemic status.

Discussion

The findings from the literature reveal a complex, landscape in which the question “Does AI know things?” is not settled — instead, it is moderated by competing epistemological criteria, differing theoretical commitments, and practical constraints. One of the strongest emergent themes is the tension between reliabilist / externalist accounts of knowledge vs internalist / justification-focused perspectives. Many works suggest that AI systems often satisfy conditions of reliability and produce outputs that align with truth (insofar as they are validated against ground truth or statistical benchmarks), but whether they meet internalist demands — such as that a knower has access to or can reflect upon justifications — remains controversial or unresolved.

Another major issue is opacity and interpretability. As AI systems become more complex (especially deep learning, large language models, etc.), their decision processes are frequently opaque. Literature like Connecting Ethics and Epistemology of AI argues for “glass-box” epistemology — that is, AI systems should be made as inspectable as possible to allow both experts and non-experts to assess their reasoning processes. Without interpretability, AI might be reliable in many cases but fall short of what internalism often demands: transparency about why something is believed or said to be known.

Epistemic agency also arises as a central concern. Several works (e.g. AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications) argue that AI changes the epistemic environment in which human agents operate, affecting belief formation and revision. AI may make belief-revision harder (if humans defer too much), or reduce critical reflection because outputs are simply accepted. Thus, even if AI doesn’t “know” in itself, its role as mediator or shaper of human knowledge practices has important epistemological consequences.

Another dimension is the ethical / normative layer that many recent works show is inseparable from epistemological assessment. Reliability, truth, justification are not only technical or philosophical metrics; they also implicate fairness, accountability, bias, social justice, and the potential for epistemic injustice. For example, Why Reliabilism Is Not Enough points out that in high-stakes contexts, a purely reliabilist justification may fail morally if the processes are opaque or if their outputs have unjust consequences. Similarly, literature like

Connecting Ethics and Epistemology of AI insists that ethical, social, and normative values must be embedded in the design, implementation, and assessment of AI systems.

There is also an indication of contextual variation and cultural norms affecting what counts as knowledge and epistemic justification. One Indonesian paper Epistemologi Kecerdasan Buatan (AI) dan Pentingnya Ilmu Etika dalam Pendidikan Interdisipliner shows that interdisciplinary education and locally rooted philosophical reflection are important. It suggests that epistemological norms might vary by cultural, educational, and institutional contexts, and that what is taken for granted in one epistemic tradition may not be in another.

From these tensions and findings, several implications follow. First, if knowledge in AI is to be acknowledged in anything like the traditional sense (belief, truth, justification), then design of AI systems must prioritize interpretability, transparency, and ways for human agents to inspect or understand the reasons behind outputs. Second, purely externalist, reliability-based accounts (while powerful) may not suffice in all contexts, especially in morally or socially sensitive domains. Third, epistemic agency must be reclaimed or preserved in human users: for instance, through training, institutional oversight, or regulatory frameworks ensuring that reliance on AI doesn't erode human ability to question and verify.

Finally, the literature also reveals gaps and open questions. For example, there is comparatively little work that examines what epistemic beliefs or belief-like states (if any) AI might have or be said to have, in virtue of their architecture. Also, although many scholars write about AI's influence on human belief formation, fewer examine whether AI itself can have metacognitive capacities (e.g., self-awareness of its own outputs, error detection) in ways relevant to epistemology. Another gap is empirical work: many philosophical claims about justification, truth, reliability are not always supported by empirical studies of how humans interact with AI, perceive AI's outputs, or trust them under varied conditions.

CONCLUSION

The review of epistemological literature reveals that while AI demonstrates features resembling human knowledge—such as reliability, truth approximation, and alignment with human expectations—it still falls short in crucial areas like belief, justificatory transparency, interpretability, and epistemic agency. Scholars generally agree that AI's reliability and superior task performance do not equate to full knowledge in the classical philosophical sense. This discussion underscores the deep interconnection between ethics and epistemology, emphasizing a "glass-box epistemology" that values transparency, inclusivity, and inspectability throughout AI's design, implementation, and evaluation. Ethical dimensions such as fairness, accountability, and social impact are essential for determining AI's legitimacy as a knowledge source. Ultimately, the question of whether AI "knows" cannot be answered simply but requires contextual nuance, clear epistemic criteria, and ethically grounded design practices to define the thresholds at which AI may contribute meaningfully to human knowledge.

REFERENCES

Alvarado, R. (2023). *AI as an Epistemic Technology. Science and Engineering Ethics, 29*(5), 1-30.

"Can We Trust Artificial Intelligence?" (2025). *Philosophy & Technology, 38*(10), Article 10. https://doi.org/10.1007/s13347-024-00820-1

Coeckelbergh, M. (2025). *AI and epistemic agency: How AI influences belief revision and its normative implications. Social Epistemology*. Advance online publication. https://doi.org/10.1080/02691728.2025.2466164

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion, 99*, Article 101896. https://doi.org/10.1016/j.inffus.2023.101896

Duran, M., & Jebeile, S. (2022). *Automating Epistemology: How AI Reconfigures Truth, Authority, and Verification. AI & Society.* (Note: approximate citation, check exact volume/page)

Ganascia, J.-G. (2010). Epistemology of AI revisited in the light of the philosophy of information. *Knowledge in Society,* *23*(1-2), 57-73. https://doi.org/10.1007/s12130-010-9101-0 (CoLab)

Heersmink, R. (2017). A virtue epistemology of the Internet: Search engines, intellectual virtues and education. *Social Epistemology, 32*(1), 1-12. https://doi.org/10.1080/02691728.2017.1383530 (Taylor & Francis Online)

Kozachenko, N. (n.d.). Artificial intelligence and academic integrity in the context of virtue epistemology. *Actual Problems of Mind*. (Check publication date.)

Ortmann. (2025). Automating epistemology: How AI reconfigures truth, authority, and verification. *AI & Society*. Advance online publication. https://doi.org/10.1007/s00146-025-02560-y

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*(3), 441-459.

Robertson, I. (2025). AI, trust and reliability. *Philosophy & Technology, 38*(3), Article 94. https://doi.org/10.1007/s13347-025-00924-2

Russo, F., Schliesser, E., & Wagemans, J. H. M. (2022). *Connecting Ethics and Epistemology of AI. AI & Society, 39*(1), 1585–1603. https://doi.org/10.1007/s00146-022-01617-6

Simion, M., & Kelp, C. (2023). Trustworthy Artificial Intelligence. *Asian Journal of Philosophy, 2*(1), Article 8. https://doi.org/10.1007/s44204-023-00063-5

(“How navigation systems transform epistemic virtues: Knowledge, issues and solutions.”) (2019). *Cognitive Systems Research, 56*, 36-49. https://doi.org/10.1016/j.cogsys.2019.03.004 (ScienceDirect)