

INSIGHTS

AI, Gettier, and the Illusion of Knowing

CHHEM Kieth Rethy¹, MD, PhD (Edu), and PhD (His)

¹ *CHHEM Kieth Rethy is a Distinguished Professor of AI and Humanities, Center for Buddhist AI Ethics, Cambodia University of Science and Technology.*

Artificial intelligence has made answers extraordinarily easy to obtain. Questions that once required a visit to a library, consultation with an expert, or hours spent searching through sources can now receive polished answers in seconds. They may be clearly written, confidently expressed, supported by references, and sometimes entirely correct. But does having the correct answer mean that we know?

But does having the correct answer mean that we know?

The question predates artificial intelligence by centuries. It acquired a particularly important modern formulation in 1963, when the American philosopher Edmund Gettier challenged the traditional account of knowledge as justified true belief. His short paper exposed a difficulty that has become unexpectedly relevant in the age of generative AI. We may arrive at a true conclusion, apparently supported by evidence, and still hesitate to call it knowledge if the conclusion happens to be true only by accident.

This is the aspect of Gettier that matters most for AI: the right answer for the wrong reasons.

Knowledge has traditionally been associated with three conditions. I believe a proposition; the proposition is true; and I have good reasons for believing it. The attraction of this formulation is obvious. Belief alone is merely conviction. Truth without belief is something I do not know. Justification connects what I believe with reasons for believing it.

Gettier showed where the difficulty lies. Truth and justification can apparently coexist while being connected in the wrong way.

Suppose I ask an AI system who currently leads a particular company. It answers, "Alex Rivera," and supplies what appear to be credible sources. I check independently and discover that Alex Rivera is indeed the CEO. I now have a belief, the belief is true, and the sources appear to justify it.

On closer examination, however, the sources turn out to concern another person called Alex Rivera, perhaps an executive working for another company. Purely by coincidence, the current CEO has exactly the same name.

The answer is correct, but the evidence did not lead reliably to it. Luck entered somewhere between the evidence and the conclusion. This is the basic difficulty exposed by Gettier (1963).

Generative AI makes the problem unusually visible because a language model does not "know" in the ordinary human sense of the word. It produces language from computational relationships learned across enormous quantities of data. The result may be fluent and convincing, but fluency

itself provides no warrant for truth. Bender et al. (2021) warned against confusing sophisticated manipulation of linguistic form with genuine understanding.

We should therefore be careful about treating plausibility, truth, and knowledge as though they were interchangeable. Something may sound right without being true. Something may even be true without our having adequate grounds for believing it. Knowledge requires a defensible connection between a claim and the evidence offered in its support.

Human judgment enters precisely at this point.

Goldman (1979) approached justification by paying closer attention to the reliability of the processes through which beliefs are formed. Applied to AI, this shifts our attention away from the final answer alone. A correct output is reassuring, but another question immediately follows: did a sufficiently reliable process produce it?

The distinction becomes consequential in medicine. Imagine an AI system analyzing a patient's clinical data and predicting coronary artery disease. Coronary angiography later confirms the diagnosis. At first sight, the algorithm has performed successfully.

Suppose, however, that further investigation shows that its prediction depended heavily on an irrelevant variable that happened to correlate with coronary disease in the training dataset. The patient had the disease, and the algorithm predicted it correctly. Yet the basis of that prediction was not medically defensible. The concern is no longer whether this particular answer happened to be right. We need to know whether the same method deserves confidence when applied to the next patient.

This is why interpretability matters in high-stakes settings. Rudin (2019) has argued that interpretable models should be preferred in such circumstances rather than relying unquestioningly on black-box systems and attempting to explain their outputs afterwards. Correctness matters, but so does the path by which correctness is reached.

The historian encounters a remarkably similar problem.

Suppose AI identifies an archaeological structure as a monastery. It cites inscriptions, architectural features, and comparisons with other sites. Later, closer examination reveals that the inscriptions actually refer to another structure. Archaeological investigation nevertheless establishes that the building was indeed a monastery.

Once again, the conclusion survives while its original justification collapses.

The stone is evidence. "Monastery" is an interpretation. Knowledge emerges from the justified relationship between the two.

We should resist the temptation, however, to regard this as a problem created by machines. Human beings have always reasoned imperfectly. We mistake correlation for causation, generalize from inadequate evidence, defer too readily to authority, attack the person instead of the argument, and preferentially notice information that confirms what we already believe. Logical fallacies existed long before computers.

What technology changes is their reach.

Social media demonstrated this dramatically. In 2015, Umberto Eco provocatively criticized the way social networks had lowered traditional barriers between private opinion and public discourse. Whatever one thinks of the severity of his language, the epistemological issue is worth considering. An unsupported assertion that might once have circulated among a handful of people could suddenly reach millions. Repetition then acquired the appearance of corroboration. A claim could become familiar without becoming better supported.

Generative AI changes the scale again. It does not merely transmit human reasoning. It can reproduce an argument, reorganize it, improve its grammar, supply persuasive transitions, and generate innumerable variations of it. When the original reasoning is weak, linguistic refinement does not repair the weakness. It may simply make the weakness harder to see.

This may be one of the more subtle epistemic dangers of AI. Bad human reasoning is sometimes recognizable because it is badly expressed. Once transformed into elegant prose, furnished with technical vocabulary, and arranged into an apparently coherent argument, the same fallacy can acquire an authority it never deserved.

Yet the capacity of AI to reformulate arguments need not work only in this direction. Used critically, the same technology can help expose hidden assumptions, identify possible fallacies, formulate objections, or place competing explanations beside one another. The decisive intellectual work remains with the person using it: examining evidence, checking sources, and deciding whether a conclusion is justified.

This gives a more precise meaning to the familiar instruction to "verify the AI." What exactly should we verify?

Checking whether a statement corresponds to the facts tests its truth. That is necessary, but it is not sufficient. We must also examine whether the evidence cited actually supports the statement. A

reference may exist without supporting the claim attributed to it. A correlation may be genuine without establishing causation. Two people may share a name; two studies may address subtly different questions; two inscriptions may belong to different monuments or historical contexts.

The responsible AI user therefore cannot remain merely a consumer of answers. The task is to examine claims. What evidence supports this conclusion? Does the source actually say what the AI claims it says? What assumptions connect the evidence to the interpretation? Could another explanation account for the same facts? Most importantly, could I be looking at the right answer for the wrong reasons?

These are epistemological questions rather than sophisticated prompting techniques.

Education in the AI era should reflect that distinction. Teaching students how to obtain better outputs is useful, but it addresses only part of the problem. Students also need to learn how to doubt a persuasive answer intelligently. They should be able to ask what they believe, whether it is true, what justifies that belief, and whether the evidence genuinely warrants the conclusion.

That is the difference between AI literacy and epistemic literacy.

The future of knowledge will not depend simply on constructing machines capable of answering more questions. It will depend equally on educating human beings who know when an answer deserves to be believed.

AI can provide answers. We still have to know why we should believe them.

The wider lesson extends beyond artificial intelligence itself:

Social media democratized expression. AI democratizes the production of apparently reasoned answers. Neither automatically democratizes knowledge.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency(pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6), 121–123. <https://doi.org/10.1093/analys/23.6.121>
- Goldman, A. I. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and knowledge* (pp. 1–23). D. Reidel. https://doi.org/10.1007/978-94-009-9493-5_1
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

* Edited with ChatGPT for clarity; all views expressed are solely those of the author.

* The opinions expressed are the authors' own and do not reflect the views of AI Forum.