THE GENERATION OF KNOWLEDGE · TRUST, QUALITY AND THE FUTURE

Machine-Generated Knowledge: Can AI Know Anything?

Whether artificial intelligence systems can be said to know anything, examined through the classic conditions for knowledge, the debate over machine understanding, the problem of opacity, the evidence on how reliable AI outputs are, and what the answer means for people who use machine-generated claims.

ModQu research article · 2,674 words · 25 references (APA 7th edition)

People now ask machines questions and receive fluent answers on almost any subject. Artificial intelligence (AI) systems summarise research, diagnose diseases from images, predict the structures of proteins and write reports. It is natural to say that such systems "know" things: that a chatbot knows the capital of Canada or that an image classifier knows what a tumour looks like. But is this more than a figure of speech? Can a machine know anything, in the sense in which people know things? The question is not only philosophical. How it is answered affects how far people should rely on machine-generated claims, and who is responsible when they are wrong. This article examines the question through the classic conditions for knowledge, the debate over machine understanding and the evidence on how reliable AI systems are.

What knowledge requires

Philosophers have long analysed knowledge as **justified true belief**. To know something, a person must believe it, it must be true, and they must have good reasons for believing it. Gettier (1963) showed that these three conditions are not quite enough. He described cases in which a person has a justified true belief that is true only by luck, and in which we would not say they know. A great deal of later philosophy tried to identify the missing ingredient.

One influential answer was **reliabilism**. Goldman (1979) proposed that a belief is justified if it is produced by a reliable process, one that tends to produce true beliefs rather than false ones. On this view, what matters is not whether the believer can state their reasons but whether the process that produced the belief is generally trustworthy. Perception, memory and sound reasoning are reliable processes; wishful thinking and guessing are not.

Reliabilism is attractive in discussions of AI because it does not require the knower to have conscious reasons. If a process is reliable, the beliefs it produces may count as knowledge, whether the process takes place in a brain or in a machine. Yet other philosophers have argued that knowledge requires more than reliability. Sosa (2007) argued that knowledge is belief that is true because of the knower's competence: the success must be creditable to the knower's own abilities, not merely the product of a process that happens to work. Grimm (2006) argued that **understanding**, grasping how things fit together and why they are as they are, is a related but distinct achievement. A person can know that a fact is true without understanding why, and understanding is often what makes knowledge useful.

Tests for knowledge, applied to a machine

Conditions philosophers have proposed for knowledge, and the open question each raises for AI systems

Condition	What it requires	The question for AI
Truth	The claim must be true	AI output is often true, but fabrications are delivered with the same fluency
Belief or commitment	The knower must hold the claim	Does a system that predicts text hold any claim at all?
Reliable process	The claim comes from a process that tends to produce truths	Can be measured: some systems are well calibrated in narrow domains
Understanding	The knower grasps why it is true and how it connects	Contested: fluent explanation may not reflect a model of the world
Accountability	Someone can answer for the claim and correct it	Rests with the people who build, use and check the system

Based on Gettier (1963), Goldman (1979), Grimm (2006) and Mitchell and Krakauer (2023).

Figure 1. Conditions for knowledge and the questions each raises for AI systems. Based on Gettier (1963), Goldman (1979), Grimm (2006) and Mitchell and Krakauer (2023); diagram by ModQu.

Can a machine believe?

Before asking whether an AI system's beliefs are justified, one must ask whether it has beliefs at all. Turing (1950) famously proposed replacing the question "Can machines think?" with a practical test: whether a machine could hold a conversation in which a human judge could not reliably tell it apart from a person. He predicted that by the end of the twentieth century, people would speak of machines thinking without expecting to be contradicted.

Searle (1980) argued that passing such a test would not show that a machine understands anything. In his **Chinese room** thought experiment, a person who does not speak Chinese sits in a room following a rule book that tells them which Chinese symbols to send out in response to the symbols passed in. To those outside, the room appears to understand Chinese. Yet the person inside understands nothing. Searle concluded that manipulating symbols according to rules, which is what computers do, is not sufficient for understanding or meaning.

The debate has been revived by modern **large language models**, which are trained on vast amounts of text to predict the next word in a sequence. Bender and Koller (2020) argued that a system trained only on the form of language, without any connection to the things words refer to, cannot learn meaning. They compared it to someone who learns to imitate conversation in a language without ever seeing the world it describes. Bender et al. (2021) described such systems as "stochastic parrots" that stitch together sequences of words according to probability, without reference to meaning.

Shanahan (2024) urged caution in the language used to describe these systems. When we say a language model "knows" or "believes" something, we risk attributing human-like mental states to a system that works in a very different way. He suggested that, strictly speaking, a basic language model answers the question "what words are likely to come next?", not "what is true?". When such a model is embedded in a larger system that consults external sources or checks its answers, the picture becomes more complicated, and some talk of belief may become more appropriate.

The debate over understanding

Mitchell and Krakauer (2023) surveyed the debate over whether large language models understand language and the world. They found the research community sharply divided. Some researchers argue that the ability of these models to solve a wide range of problems, including some they were not specifically trained on, shows a form of understanding. Others point out that models often fail in ways no person with genuine understanding would, for example by making elementary errors on problems that are slightly altered from familiar examples, which suggests that they rely on statistical associations rather than a model of how the world works.

Mitchell and Krakauer suggested that the disagreement may partly reflect different concepts of understanding. Machines may possess a kind of understanding quite different from human understanding, one that allows them to exploit statistical regularities across vast amounts of data but not to form the compact causal models that humans use to reason about new situations. If so, the question "Does AI understand?" has no simple answer. It depends on what kind of understanding is meant and what tasks it must support.

Floridi and Chiriatti (2020) tested an early large language model on questions of mathematics, meaning and ethics. They found that it could produce text that was fluent and often impressive but that it failed on simple questions requiring genuine comprehension, and concluded that such systems should be understood as tools that produce text, not as agents that understand it.

How reliable is machine-generated content?

Whatever the philosophical answer, the practical question is how far machine-generated claims can be relied upon. Here, the evidence is mixed and depends greatly on the type of system and the task.

Some systems are remarkably reliable in narrow domains. Jumper et al. (2021) showed that AlphaFold predicted the structures of proteins with accuracy often comparable to experimental methods. In such cases, the system has been trained on high-quality data and is tested against independent measurements, and its reliability can be measured precisely.

General-purpose language models are less dependable. Ji et al. (2023) reviewed research on **hallucination**, the tendency of language generation systems to produce content that is fluent but false or unsupported by any source. Hallucinations include invented facts, misattributed quotations and references to publications that do not exist. Because they are expressed with the same fluency as accurate content, they are hard to detect without independent checking.

Lin et al. (2022) designed a test of truthfulness, made up of questions that some people answer falsely because of common misconceptions. The best language model tested was truthful on just over half of the questions, while people were truthful on the large majority. Strikingly, larger models were often less truthful than smaller ones, because they had learned to reproduce popular misconceptions found in their training data more faithfully.

There is some evidence that models can estimate their own reliability. Kadavath et al. (2022) found that large language models were often reasonably well **calibrated** when asked to judge whether their own answers were correct: the probability they assigned to an answer being true corresponded approximately to how often such answers were actually true. This kind of self-assessment is important, since a source that knows when it might be wrong is much more useful than one that is uniformly confident. However, calibration varied across tasks and was weaker on unfamiliar kinds of questions.

Reliability also depends on how people use AI output. Research on automation has shown that people tend to over-rely on automated aids. Parasuraman and Manzey (2010) reviewed studies of **automation bias** and **complacency** in settings such as aviation and medicine. When an automated system was usually right, people checked its output less carefully and missed the errors it made, and this happened to experienced experts as well as novices. Training and instructions alone did little to prevent it. The lesson for AI is that a system which is right most of the time can still lead people into error, precisely because its usual reliability discourages the checking that would catch its mistakes.

The problem of opacity

A further difficulty concerns the transparency of AI systems. Humphreys (2009) described computational methods as **epistemically opaque**: the process by which they reach their results is too complex for any human to follow step by step. Burrell (2016) distinguished three sources of opacity in machine learning: intentional secrecy by companies, the technical illiteracy of most users, and the intrinsic complexity of models that operate with millions or billions of parameters, which makes their reasoning hard to interpret even for experts.

Opacity matters for knowledge. If a person cannot see why a system reached a conclusion, they cannot assess whether its reasons were good. Lipton (2018) noted that the term **interpretability** is used in many different senses, and that some popular methods for explaining machine learning models provide plausible stories rather than faithful accounts of how the models actually work. Rudin (2019) argued that for decisions with serious consequences, such as in criminal justice and health care, it is better to use models that are interpretable by design than to attempt to explain black-box models after the fact.

Durán and Formanek (2018) proposed a way of trusting opaque systems without understanding their inner workings. Under what they called **computational reliabilism**, trust in a computer's results is justified when there is good evidence that the process is reliable, for example through verification and validation of the system, a history of successful use and the endorsement of experts in the relevant field. This is similar to the way people trust instruments such as thermometers or microscopes without understanding every detail of how they work.

Knowledge without a knower?

Some philosophers have argued that knowledge need not reside in a mind at all. Popper (1972) distinguished between knowledge in the subjective sense, the beliefs and expectations held by individuals, and knowledge in the objective sense: theories, arguments and problems as they exist in books, journals and libraries, independent of anyone who happens to believe them. Objective knowledge, he argued, can be criticised, tested and improved, and it can contain consequences that no one has yet noticed. On this view, a database or a model might contain knowledge in the objective sense, even if no one, human or machine, believes it.

Clark and Chalmers (1998) went further with their idea of the **extended mind**. They described a man with memory loss who relies on a notebook to record addresses he would otherwise forget. If the notebook is reliably available, trusted and consulted as readily as memory, they argued, it is part of his cognitive system, and the information in it counts among his beliefs. Applied to AI, this suggests that the right unit for asking about knowledge may be not the machine alone but the combined system of a person and the tools they rely on. Whether that system knows depends on how reliably the person uses the tool, and how well they understand its limits.

AI as testimony

Most of what people know comes from others. Hardwig (1985) argued that rational belief often depends on trusting experts whose reasons we cannot assess ourselves. Lackey (2008) argued that what matters in learning from testimony is not mainly the speaker's beliefs but whether their statements are reliable sources of truth. On this view, a person may come to know something by reading a reliable book, even if the author was mistaken in their private beliefs, so long as the statement in the book is true and reliably produced.

This suggests one way of thinking about AI. Even if an AI system does not know anything in the full sense, people may come to know things through it, just as they come to know things through reference works, instruments and other people. Whether they do depends on the reliability of the system for the kind of question being asked, and on whether the person has good reason to believe that it is reliable. The burden of knowledge, in other words, falls on the user and the institutions that build and check the system, not on the machine alone.

There is, however, an important difference between AI and human testimony. A human expert can be held to account. They can be asked for their reasons, challenged by peers and held responsible for errors. Their reputation depends on their accuracy. AI systems cannot be held responsible in this way. Responsibility must therefore rest with the people and organisations who build, deploy and use them.

Implications for producers and users of knowledge

For organisations that produce knowledge, including survey and market research organisations, the question of whether AI can know anything has practical consequences.

First, machine-generated claims should be treated as testimony from a source of uncertain and variable reliability, not as knowledge. They should be checked against original sources, especially facts, figures and references, which language models are prone to invent.

Second, reliability should be measured, not assumed. When AI is used for a defined task, such as classifying open-ended survey responses, its performance should be compared with careful human work on a sample of cases, and the results should be reported.

Third, transparency should be preserved. Reports should state where and how AI tools were used, so that readers can judge how much weight to give the findings. Where decisions have serious consequences, interpretable methods should be preferred.

Fourth, accountability must remain with people. An organisation that publishes findings is responsible for them, whatever tools it used to produce them. "The AI said so" is not a justification.

Fifth, human expertise must be maintained. The research on automation bias shows that people who rely on automated aids tend to check them less and to lose the skills they would need to notice errors. An organisation that hands all of its coding, analysis or drafting to AI may find that, after a few years, no one on its staff can tell whether the output is right. Keeping experienced people involved in the work, and giving newer staff the chance to do tasks themselves before they automate them, protects the organisation's capacity to judge what its tools produce. In the extended-mind sense described above, the knowledge belongs to the combined system of people and tools, and that system is only as reliable as the people's ability to recognise when the tools have gone wrong.

Conclusion

Can AI know anything? On the classic analysis, knowledge requires belief, truth and justification, and perhaps understanding and the competence of the knower. AI systems produce many true outputs, and some, in narrow domains, are highly reliable. But whether they hold beliefs or understand anything remains contested, and general-purpose language models produce falsehoods with the same fluency as truths. A useful way to think about AI is as a source of testimony: people can come to know things through it when it is reliable for the task and when they have good reason to trust it. That reason must come from measurement, validation and checking, and responsibility for the resulting knowledge must remain with the people and institutions who use these systems. Machines can help generate knowledge. Whether they can know is a question that, for now, matters less than whether we have checked what they tell us.

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

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

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

Grimm, S. R. (2006). Is understanding a species of knowledge? *The British Journal for the Philosophy of Science*, *57*(3), 515–535. https://doi.org/10.1093/bjps/axl015

Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, *82*(7), 335–349. https://doi.org/10.2307/2026523

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, *169*(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). *Language models (mostly) know what they know* (arXiv:2207.05221). arXiv. https://doi.org/10.48550/arXiv.2207.05221

Lackey, J. (2008). *Learning from words: Testimony as a source of knowledge*. Oxford University Press.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Clarendon Press.

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

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67*(2), 68–79. https://doi.org/10.1145/3624724

Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. 1). Oxford University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433