AI and human intelligence are drastically different—here’s how

Walter QuattrociocchiMay 2026

Putting humans and LLMs head-to-head in classic tests of judgment from human psychology underscores the differences between them

When you walk into your doctor’s office, you assume something so basic that it barely needs articulation: the doctor has touched a body before. They have studied anatomy, seen organs, and learned the difference between pain that radiates and pain that pulses. They have developed this knowledge, you believe, not only through reading but after years of hands-on experience and training.

Now imagine discovering that this doctor has never encountered a body at all. Instead they have merely read millions of patient reports and learned, in exquisite detail, how a diagnosis typically “sounds.” Their explanations might still feel persuasive, even comforting. The cadence would be right, the vocabulary impeccable, the formulations reassuringly familiar. Yet the moment you learned what their knowledge was actually based on—patterns in text rather than contact with the world—something essential would dissolve.

Every day many people turn to tools such as OpenAI’s ChatGPT for medical advice, legal guidance, psychological insight, educational tutoring, or judgments about what is and isn’t true. And on some level they know these large language models (LLMs) are imitating an understanding of the world they don’t actually have—even if their fluency can make that easy to forget.

But is an LLM’s reasoning anything like human judgment, or is it merely generating the linguistic silhouette of reasoning? As a scientist who studies human judgment and the dynamics of information, I recently set out with my colleagues to address this surprisingly underexplored question. We compared how LLMs and people responded when asked to make judgments across a handful of tests that have been studied for decades in psychology and neuroscience. We didn’t expect these systems to “think” like people, but we believed it would be valuable to understand how such tools differ from humans to help people evaluate how and when to use them.

In one experiment, we presented 50 people and six LLMs with a variety of news sources, then asked them to rate each source’s

credibility and justify their rating. Past research shows that when a person encounters a questionable headline, several things typically happen. First, the person checks the headline against what they already know about the world to decide whether it fits with basic facts, past events or personal experience. Second, the reader brings in expectations about the source, such as whether it is an outlet with a history of careful reporting or one known for exaggeration or bias. Third, the person considers whether the claim makes sense as part of a broader chain of events, whether it could realistically have happened and whether it aligns with the way similar situations usually unfold.

> LLMs can often match human responses but for reasons that bear no resemblance to human reasoning.

LLMs cannot carry out these steps. To see what they do instead, we asked leading models to evaluate the reliability of news headlines by following a specific procedure. We instructed the LLMs to state the criteria they were using to determine credibility and to justify their final judgment. We observed that even when models reached conclusions similar to those of human participants, their justifications consistently reflected patterns drawn from language (such as how often a particular combination of words coincided and in what contexts) rather than references to external facts, prior events or experience, which were the factors that humans considered.

In other experiments, we compared humans' and LLMs' reasoning around moral dilemmas. When humans think about morality, they draw on norms, social expectations, emotional responses, and culturally shaped intuitions about harm and fairness. As one example, when people evaluate morality, they often use causal reasoning: They consider how one event leads to another, why timing matters and how things might have turned out differently if something had changed along the way. People imagine different situations through counterfactuals in which they ask, "What if this circumstance had been different?"

We found that a language model can reproduce this form of deliberation fairly well. The model provides statements that mirror the vocabulary of care, duty or rights. It will present causal language based on patterns in language, including "if-then" counterfactuals. But it's important to note that the model is not imagining anything or engaging in any deliberation; it is just reproducing patterns in people's speech or writing about these counterfactuals. The result can sound like causal reasoning, but the process behind it is pattern completion, not an understanding of how events produce outcomes in the world.

Across all the tasks we have studied, a consistent pattern emerges. LLMs can often match human responses but for reasons that bear no resemblance to human reasoning. Where a human judges, a model correlates. Where a human evaluates, a model predicts. Where a human engages with the world, a model engages with a distribution of words. Their architecture makes them extraordinarily good at reproducing patterns found in text. It does not give them access to the world those words refer to.

And yet, because human judgments are also expressed through language, the model's answers often end up resembling human answers on the surface. This gap between what models seem to be doing and what they are in fact doing is what my colleagues and I call epistemia: a situation when the simulation of knowledge becomes indistinguishable, to the observer, from knowledge itself. Epistemia is a flaw in people's interpretation of these models in which linguistic plausibility is taken as a surrogate for truth. This error happens because the model is fluent, and fluency is something human readers are primed to trust.

The danger here is subtle. It is not primarily that models are often wrong—people can be, too. The deeper issue is that the model cannot know when it is "hallucinating," because it cannot represent truth in the first place. It cannot form beliefs, revise them or check its output against the world. It cannot distinguish a reliable claim from an unreliable one except by analogy to prior linguistic patterns. In short, it cannot do what judgment is fundamentally for.

People are already using these systems in contexts in which it is necessary to distinguish between plausibility and truth, such as law, medicine and psychology. A model can generate a paragraph that sounds like a diagnosis, a legal analysis or a moral argument. But sound is not substance. The simulation is not the thing simulated.

None of this implies that LLMs should be rejected. They are extraordinarily powerful tools when used as what they are: engines of linguistic automation, not engines of understanding. They excel at drafting, summarizing, recombining and exploring ideas. But when we ask them to judge, we unintentionally redefine judgment—shifting it from a relation between a mind and the world to one between a prompt and a probability distribution.

What should a reader do with this knowledge? Do not fear these systems but instead seek a clearer understanding of what they can and cannot do. Remember that smoothness is not insight, and eloquence is not evidence of understanding. Treat large language models as sophisticated linguistic instruments that require human oversight precisely because they lack access to the domain that judgment ultimately depends on: the world itself.

Are you a scientist who specializes in neuroscience, cognitive science or psychology? And have you read a recent peer-reviewed paper that you would like to write about for Mind Matters? Please send suggestions to Scientific American*'s Mind Matters editor Daisy Yuhas at* dyuhas@sciam.com.