

AI Could Worsen a Common Cognitive Trap

Tania Lombrozo

One of humanity's most remarkable cultural achievements is the division of explanatory labor. When my daughter came down with a fever, I solved the problem by consulting a pediatrician, who shared a likely diagnosis of the flu. Unlike the doctor, I wasn't able to tell my daughter much about the mechanism by which the flu leads to fever, or exactly why it was so important for her to keep drinking fluids. But the explanation I acquired—roughly “fever because flu”—was what was needed to treat her.

Society runs on this efficient type of information sharing. Just as coordinating specialized skills can streamline the fabrication of products—be it pins or cars or newspapers—coordinating specialized knowledge can facilitate the answering of questions. You can drive a car even if you don't have the expertise to manufacture one yourself, and I benefited from the pediatrician's explanation even though his understanding far exceeded my own.

We can thank this divide-and-conquer method for many advancements in medicine and science, engineering and commerce; in fact, it has helped drive the accumulation of human knowledge across communities and over time. But it also leaves us susceptible to illusions of understanding: cases in which we mistake the understanding of others for our own.

Humans' sense of whom to ask for what is fine-tuned by evolution, culture, and personal experience. Even 3-to-5-year-olds are more likely to say that you should ask a doctor rather than a mechanic about fixing a broken arm, and a mechanic rather than a doctor about fixing a flat tire. As kids mature, they develop an even deeper understanding of expertise. In [a 2004 study published in *Child Development*](#), for example, 11- and 12-year-olds tasked with finding out why tennis balls bounce better on the sidewalk than on grass were more likely to suggest asking someone who knows why bubble wrap keeps glass things from breaking (which is also relevant to physics) than someone who knows why tennis balls come in cans of three (which just happens to also involve tennis balls). Their choice of whom to ask was based on fundamental principles—not just superficial similarities.

With plenty of time and care, years of education, and a healthy appetite for learning, most of us transition from children who know little about science or other expert domains to ... adults who know little about science or other expert domains. Even those of us who acquire expertise in one area are unlikely to do so in many others. So the ability to track outside expertise is no less important for adults than it is for children. We need to know who can explain a child's fever, a wilting plant, a clause in a will, or an imperfect soufflé—and research finds that adults seem to do this pretty well. (As a population, of course, we do not do this perfectly: Plenty of people trust in bad advice from their friends, influencers, and self-proclaimed experts.) A study from Germany found that even college students who aren't science majors do a good job of guessing which kinds of scientific expertise are relevant to different scientific topics.

So far, so good. The trouble arises when our connections to other minds become a little too seamless, and we start to lose track of where our own understanding ends and another's begins. When we receive an explanation from an expert—such as “fever because flu”—we might mistakenly conclude that we've learned more than we really have.

In fact, research finds that merely having access to expert answers can lead to illusions of understanding. In one set of studies, participants read about various fictional scientific discoveries and reported how well they understood the phenomenon described in each one. For instance, some participants received the following short passage:

A May 19, 2014, study in the journal *Geology* reported the discovery of a new rock that scientists have thoroughly explained. The rock is similar to calcite, yet it glows in the absence of a light source. The authors of the study, Rittenour, Clark, and Xu, fully understand how it works; they provided a description of the remarkable appearance of the mineral and outlined future experiments.

The passage didn't actually include any explanations about why the rock glows. And yet, after reading the passage, participants thought they had some comprehension: what the authors call “a contagious sense of understanding,” because they seem to have caught it from the scientists described in the passage. When participants instead read that scientists “have not yet explained” the glowing rock, and that they “do not yet understand” how it works, there was no contagious sense of understanding; participant ratings of their own understanding nearly hit the bottom of the scale.

More recent work has shown that a contagious sense of understanding can even infect our interactions with technology. In one study, participants learned about a bank's “robo-adviser,” which could provide investment recommendations. Participants also learned that an explanation of how the robo-adviser works had been published in a scientific journal. But half were told that the explanation was available to them on the journal's website, and half were told that the journal had a subscription paywall. Those who thought the explanation was available to them reported a better understanding of how the robo-adviser works. They also reported a greater likelihood of using the robo-adviser themselves.

These results raise a pressing question. As information becomes more accessible through AI assistants, chatbots, and other technology, will these illusions of understanding become even more pronounced? Some early evidence suggests that the answer is yes. Several studies have now found that when people have information available to them through the internet, they overestimate how much they themselves know.

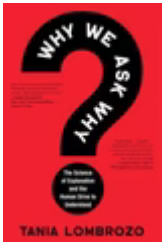
For example, in one study, participants were asked to explain how something works, such as a zipper, and half were given access to the internet to “confirm the details of the explanation.” All participants were then asked how well they could offer detailed answers to unrelated questions, such as why cloudy nights are warm. Those who had access to the internet in the first part of the study rated themselves significantly more capable of explaining in the second part—suggesting that the experience of having more knowledge at their fingertips supported the illusion that the knowledge was in their own heads.

Interactions with fast and informative chatbots will only heighten such illusions, but with a worrisome twist: The evidence so far suggests that humans aren't great at tracking who knows what when it comes to large

language models such as ChatGPT and Claude. For instance, when people in a study learned about an LLM that could correctly answer a question about economics, they expected it to correctly solve a basic algebra problem—but they were wrong. And when they learned that an LLM could answer a question about moral philosophy, they thought it could identify which character in a scenario had done something immoral—but they were wrong again. Participants almost certainly applied assumptions about how knowledge clusters in human minds to the AI system, and in both cases, this led them astray.

The kinds of errors that AI systems make are likely to change as AI develops, and human intuition about AI will probably evolve alongside these changes. As we expand our community of explainers beyond human experts to chatbots and other forms of AI, we create new opportunities for human understanding. But we also introduce an increased potential for illusion and error.

This article was adapted from Tania Lombrozo's new book, [*Why We Ask Why: The Science of Explanation and the Human Drive to Understand*](#).



By Lombrozo, Tania

When you buy a book using a link on this page, we receive a commission. Thank you for supporting The Atlantic.