Artificial intelligence shares our confidence bias, research reveals

November 29 2023

In experiments, an artificial-intelligence model was able to distinguish noisy images of numbers similar to the way that humans do. Credit: RIKEN Center for Brain Science

The pronounced positive confidence bias that is a characteristic and apparently irrational trait of human decision-making has been replicated and dissected using an artificial intelligence (AI) model. The unnerving [result](link), published in *Nature Communications* by a RIKEN-led team, reveals that our inflated sense of confidence might stem from subtle observational cues.

When we glimpse something familiar, we're often able to instantly and confidently conclude a match to that familiar object, even if the balance of evidence doesn't support such a high level of confidence. This dissociation between decision-making and confidence level has long perplexed researchers, as it suggests that while we are capable of highly rational [decision-making](link), our sense of confidence about our decisions can be quite irrational.

"There's been a tension between theory, which bluntly assumes that humans are rational, and the [empirical data](link), which clearly shows that this is not always the case," notes Hakwan Lau of the RIKEN Center for Brain Science.

This generally occurs when the image is unclear, or noisy. Mathematically, the noisiness of an image can be calculated from its deviation from the clear image by a metric called the signal-to-[noise](link) ratio.

But this is where things start to get weird. "If I make an image both more salient and more noisy, but keep the overall signal-to-[noise ratio](link) the same, we somehow become more confident in believing we know what we're seeing, even though we aren't seeing any better," says Lau. "It turns out that the structure of the noise, which is usually assumed to be random, actually matters a lot."

Now, Lau and co-workers have investigated the effect of different types

of noise on decision confidence using an AI model that specifically reported on confidence.

"One always wonders what an AI model is doing," says Lau. "But the great thing with an AI model is that once it learns, unlike with the human brain, we can 'dissect' the model to understand it better."

Surprisingly, the AI models showed exactly the same confidence biases as humans. This seems to be exactly what it is supposed to do, as Lau points out.

"The model is rational in a sense that it learns from the noise structure of natural images, which is different from the standard noise type assumed in signal-processing models," says Lau. "It's the learning of the statistical properties of natural images that lead these models—and presumably our brains too—to have these apparent biases."

More information: Taylor W. Webb et al, Natural statistics support a rational account of confidence biases, *Nature Communications* (2023). DOI: 10.1038/s41467-023-39737-2

Provided by RIKEN