Hey! What's New? 2026-34

AI Works Better When It's a Little Bit Human

Here's an article about AI that doesn't pertain directly to the accounting profession but is still of considerable interest to those looking to use it. An article in the *Chicago Booth Review*, written by Monika Brown, says that "when people think about how machines process information, they tend to think of a cold, rational operation. But training artificial intelligence with a measure of human misperception might be the key to making it smarter and cheaper."

She writes that Princeton PhD student Sijia Liu, Stanford PhD student Niklas Muennighoff, and Chicago Booth's Kawin Ethayarajh looked into the difference between the two broad approaches AI developers currently use to align their models after pretraining is complete. "They find that the better-performing class of methods serendipitously reflects human biases about probabilities – an insight that explains why the industry's priciest training techniques work so well. This finding allowed the researchers to create an approach that they contend matches the quality of the expensive method at a fraction of the cost."

Brown notes that training a state-of-the-art language model can cost more in a week than a small startup spends in a year. "A growing fraction of that expense comes from alignment – a process that trains the model using feedback signals on its outputs – after they've absorbed the huge datasets that form the foundation of their learning. Alignment is what shapes raw model capabilities into outputs that are actually useful and appropriate, whether in the context of safety – can we prevent the model from being used for hacking? – or simply in the context of correctness – is the model doing the right mathematical reasoning?"

Alignment can be broken down into two broad types, she says: "offline, in which the model learns from a fixed dataset that teaches it how to behave, and online. In online alignment training, the model generates outputs, receives automated feedback from a scoring system, generates more outputs using that feedback, and repeats. It's like training chefs by having them cook dish after dish from scratch and critiquing them after each one instead of simply giving them cookbooks to read."

Online alignment methods are slow and costly but yield better results than those using the offline approach, the researchers found. "Conventional wisdom credits the constant flow of fresh data that the model generates during training, but Liu, Muennighoff, and Ethayarajh's research suggests online alignment succeeds at least

partly because it accidentally forces language models to learn in a way that mirrors human psychology."

Humans systematically distort probability, they found. "Someone might overestimate the chance of winning a massive lottery jackpot while simultaneously underestimating the chance of earning a modest return on an index fund. Behavioral economists call this 'probability weighting,' a central concept in the behavioral framework known as prospect theory. This theory explains how we overestimate extreme, rare outcomes and underestimate common, highly probable ones."

"Crucially, this distortion is systematic – not random – and the shape of the distortion is consistent across humans," Ethayarajh says. "If, instead of thinking about the amount of money that might be gained, we think about the amount of information that might be gained, we can frame alignment through the lens of behavioral economics."

Aligning a model – training it to give answers that people judge as helpful, accurate and appropriate – with pricier, continuously generated, online data typically produces better performance than with cheaper, fixed "offline" data. But, notes Brown, "the researchers' approach, which they call "humanline" because it adjusts training to reflect human biases, can make offline alignment comparable to the online method. The results hold regardless of the training algorithm used."

Applying that frame, says Brown, "the researchers find evidence that online alignment distorts probabilities just like humans do. Much the way winning a jackpot occupies an outsize position in our minds, so do certain model outputs play an outsize role during training."

Liu, Muennighoff, and Ethayarajh's key insight is that "if the online alignment method succeeds by accidentally mimicking human perceptual biases, those biases could be deliberately incorporated into any training approach. The key to effective alignment may be as much about how well the method matches human perception as where its data come from."

Learn more about where this research led and how it can best be used at AI Works Better When It's a Little Bit Human | Chicago Booth Review.