https://pluralistic.net/2024/04/23/maximal-plausibility/#reverse-centaurs

"Humans in the loop" must detect the hardest-to-spot errors, at superhuman speed

Cory Doctorow -- 23-04-2024

If AI has a future (a big if), it will have to be economically viable. An industry can't spend 1,700% more on Nvidia chips than it earns indefinitely – not even with Nvidia being a principle investor in its largest customers:

https://news.ycombinator.com/item?id=39883571

A company that pays 0.36-1 cent/query for electricity and (scarce, fresh) water can't indefinitely give those queries away by the millions to people who are expected to revise those queries dozens of times before eliciting the perfect botshit rendition of "instructions for removing a grilled cheese sandwich from a VCR in the style of the King James Bible":

https://www.semianalysis.com/p/the-inference-cost-of-search-disruption

Eventually, the industry will have to uncover some mix of applications that will cover its operating costs, if only to keep the lights on in the face of investor disillusionment (this isn't optional – investor disillusionment is an inevitable part of every bubble).

Now, there are *lots* of low-stakes applications for AI that can run just fine on the current AI technology, despite its many – and seemingly inescapable – errors ("hallucinations"). People who use AI to generate illustrations of their D&D characters engaged in epic adventures from their previous gaming session don't care about the odd extra finger. If the chatbot powering a tourist's automatic text-to-translation-to-speech phone tool gets a few words wrong, it's still much better than the alternative of speaking slowly and loudly in your own language while making emphatic hand-gestures.

There are lots of these applications, and many of the people who benefit from them would doubtless pay something for them. The problem – from an AI company's perspective – is that these aren't just low-*stakes*, they're also low-*value*. Their users would pay *something* for them, but not very much.

For AI to keep its servers on through the coming trough of disillusionment, it will have to locate *high*-value applications, too. Economically speaking, the function of low-value applications is to soak up excess capacity and produce value at the margins after the high-value applications pay the bills. Low-value applications are a side-dish, like the coach seats on an airplane whose total operating expenses are paid by the business class passengers up front. Without

the principle income from high-value applications, the servers shut down, and the low-value applications disappear:

https://locusmag.com/2023/12/commentary-cory-doctorow-what-kind-of-bubble-is-ai/

Now, there are lots of high-value applications the AI industry has identified for its products. Broadly speaking, these high-value applications share the same problem: they are all high-stakes, which means they are very sensitive to errors. Mistakes made by apps that produce code, drive cars, or identify cancerous masses on chest X-rays are extremely consequential.

Some businesses may be insensitive to those consequences. Air Canada replaced its human customer service staff with chatbots that just lied to passengers, stealing hundreds of dollars from them in the process. But the process for getting your money back after you are defrauded by Air Canada's chatbot is so onerous that only one passenger has bothered to go through it, spending ten weeks exhausting all of Air Canada's internal review mechanisms before fighting his case for weeks more at the regulator:

https://bc.ctvnews.ca/air-canada-s-chatbot-gave-a-b-c-man-the-wrong-information-now-the-airline-has-to-pay-for-the-mistake-1.6769454

There's never just one ant. If this guy was defrauded by an AC chatbot, so were hundreds or thousands of other fliers. Air Canada doesn't have to pay them back. Air Canada is tacitly asserting that, as the country's flagship carrier and near-monopolist, it is too big to fail and too big to jail, which means it's too big to *care*.

Air Canada shows that for some business customers, AI doesn't need to be able to do a worker's job in order to be a smart purchase: a chatbot can replace a worker, *fail* to their worker's job, and still save the company money on balance.

I can't predict whether the world's sociopathic monopolists are numerous and powerful enough to keep the lights on for AI companies through leases for automation systems that let them commit consequence-free free fraud by replacing workers with chatbots that serve as moral crumple-zones for furious customers:

https://www.sciencedirect.com/science/article/abs/pii/S0747563219304029

But even stipulating that this is sufficient, it's intrinsically unstable. Anything that can't go on forever eventually stops, and the mass replacement of humans with high-speed fraud software seems likely to stoke the already blazing furnace of modern antitrust:

https://www.eff.org/de/deeplinks/2021/08/party-its-1979-og-antitrust-back-baby

Of course, the AI companies have their own answer to this conundrum. A high-stakes/high-value customer can still fire workers and replace them with AI –

they just need to hire fewer, cheaper workers to supervise the AI and monitor it for "hallucinations." This is called the "human in the loop" solution.

The human in the loop story has some glaring holes. From a worker's perspective, serving as the human in the loop in a scheme that cuts wage bills through AI is a nightmare – the worst possible kind of automation.

Let's pause for a little detour through automation theory here. Automation can *augment* a worker. We can call this a "centaur" – the worker offloads a repetitive task, or one that requires a high degree of vigilance, or (worst of all) both. They're a human head on a robot body (hence "centaur"). Think of the sensor/vision system in your car that beeps if you activate your turn-signal while a car is in your blind spot. You're in charge, but you're getting a second opinion from the robot.

Likewise, consider an AI tool that double-checks a radiologist's diagnosis of your chest X-ray and suggests a second look when its assessment doesn't match the radiologist's. Again, the human is in charge, but the robot is serving as a backstop and helpmeet, using its inexhaustible robotic vigilance to augment human skill.

That's centaurs. They're the good automation. Then there's the *bad* automation: the *reverse*-centaur, when the human is used to augment the robot.

Amazon warehouse pickers stand in one place while robotic shelving units trundle up to them at speed; then, the haptic bracelets shackled around their wrists buzz at them, directing them pick up specific items and move them to a basket, while a third automation system penalizes them for taking toilet breaks or even just walking around and shaking out their limbs to avoid a repetitive strain injury. This is a robotic head using a human body – and destroying it in the process.

An AI-assisted radiologist processes *fewer* chest X-rays every day, costing their employer *more*, on top of the cost of the AI. That's not what AI companies are selling. They're offering hospitals the power to create reverse centaurs: radiologist-assisted AIs. That's what "human in the loop" means.

This is a problem for workers, but it's also a problem for their bosses (assuming those bosses actually care about correcting AI hallucinations, rather than providing a figleaf that lets them commit fraud or kill people and shift the blame to an unpunishable AI).

Humans are good at a lot of things, but they're not good at *eternal, perfect vigilance*. Writing code is hard, but performing code-review (where you check someone else's code for errors) is much harder – and it gets *even harder* if the code you're reviewing is *usually* fine, because this requires that you maintain

your vigilance for something that only occurs at rare and unpredictable intervals:

https://twitter.com/qntm/status/1773779967521780169

But for a coding shop to make the cost of an AI pencil out, the human in the loop needs to be able to process a *lot* of AI-generated code. Replacing a human with an AI doesn't produce any savings if you need to hire two more humans to take turns doing close reads of the AI's code.

This is the fatal flaw in robo-taxi schemes. The "human in the loop" who is supposed to keep the murderbot from smashing into other cars, steering into oncoming traffic, or running down pedestrians isn't a driver, they're a driving *instructor*. This is a *much* harder job than being a driver, even when the student driver you're monitoring is a human, making human mistakes at human speed. It's even harder when the student driver is a robot, making errors at computer speed:

https://pluralistic.net/2024/04/01/human-in-the-loop/#monkey-in-the-middle

This is why the doomed robo-taxi company Cruise had to deploy 1.5 skilled, high-paid human monitors to oversee each of its murderbots, while traditional taxis operate at a fraction of the cost with a single, precaratized, low-paid human driver:

https://pluralistic.net/2024/01/11/robots-stole-my-jerb/#computer-says-no

The vigilance problem is pretty fatal for the human-in-the-loop gambit, but there's another problem that is, if anything, *even more* fatal: the *kinds* of errors that AIs make.

Foundationally, AI is applied statistics. An AI company trains its AI by feeding it a *lot* of data about the real world. The program processes this data, looking for statistical correlations in that data, and makes a model of the world based on those correlations. A chatbot is a next-word-guessing program, and an AI "art" generator is a next-pixel-guessing program. They're drawing on billions of documents to find the most statistically likely way of finishing a sentence or a line of pixels in a bitmap:

https://dl.acm.org/doi/10.1145/3442188.3445922

This means that AI doesn't just make errors – it makes *subtle* errors, the kinds of errors that are the *hardest* for a human in the loop to spot, because they are the most statistically probable ways of being wrong. Sure, we *notice* the gross errors in AI output, like confidently claiming that a living human is dead:

https://www.tomsguide.com/opinion/according-to-chatgpt-im-dead

But the most common errors that AIs make are the ones we don't notice, because they're perfectly camouflaged as the truth. Think of the recurring AI

programming error that inserts a call to a nonexistent library called "huggingface-cli," which is what the library would be called if developers reliably followed naming conventions. But due to a human inconsistency, the real library has a slightly different name. The fact that AIs repeatedly inserted references to the nonexistent library opened up a vulnerability – a security researcher created a (inert) malicious library with that name and tricked numerous companies into compiling it into their code because their human reviewers missed the chatbot's (statistically indistinguishable from the truth) lie:

https://www.theregister.com/2024/03/28/ai_bots_hallucinate_software_packages/

For a driving instructor or a code reviewer overseeing a human subject, the majority of errors are comparatively easy to spot, because they're the kinds of errors that lead to inconsistent library naming – places where a human behaved erratically or irregularly. But when *reality* is irregular or erratic, the AI will make errors by presuming that things are statistically normal.

These are the hardest kinds of errors to spot. They couldn't be harder for a human to detect if they were *specifically designed* to go undetected. The human in the loop isn't just being asked to spot mistakes – they're being actively deceived. The AI isn't merely wrong, it's constructing a subtle "what's wrong with this picture"-style puzzle. Not just one such puzzle, either: millions of them, at speed, which must be solved by the human in the loop, who must remain perfectly vigilant for things that are, by definition, almost totally unnoticeable.

This is a special new torment for reverse centaurs – and a significant problem for AI companies hoping to accumulate and keep enough high-value, high-stakes customers on their books to weather the coming trough of disillusionment.

This is pretty grim, but it gets grimmer. AI companies have argued that they have a third line of business, a way to make money for their customers beyond automation's gifts to their payrolls: they claim that they can perform difficult scientific tasks at superhuman speed, producing billion-dollar insights (new materials, new drugs, new proteins) at unimaginable speed.

However, these claims – credulously amplified by the non-technical press – keep on shattering when they are tested by experts who understand the esoteric domains in which AI is said to have an unbeatable advantage. For example, Google claimed that its Deepmind AI had discovered "millions of new materials," "equivalent to nearly 800 years' worth of knowledge," constituting "an order-of-magnitude expansion in stable materials known to humanity":

https://deepmind.google/discover/blog/millions-of-new-materials-discovered-with-deep-learning/

It was a hoax. When independent material scientists reviewed representative samples of these "new materials," they concluded that "no new materials have been discovered" and that not *one* of these materials was "credible, useful and novel":

https://www.404media.co/google-says-it-discovered-millions-of-new-materials-with-ai-human-researchers/

As Brian Merchant writes, AI claims are eerily similar to "smoke and mirrors" – the dazzling reality-distortion field thrown up by 17th century magic lantern technology, which millions of people ascribed wild capabilities to, thanks to the outlandish claims of the technology's promoters:

https://www.bloodinthemachine.com/p/ai-really-is-smoke-and-mirrors

The fact that we have a four-hundred-year-old name for this phenomenon, and yet we're still falling prey to it is frankly a little depressing. And, unlucky for us, it turns out that AI therapybots can't help us with this – rather, they're apt to literally convince us to kill ourselves:

https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says