THE AGE

Explainer

AI

‘Apologies for any confusion’: Why chatbots hallucinate

Eager to please, over-confident and sometimes downright deceptive. If that sounds like the chatbot in your life, you’re not the only one. How often does artificial intelligence get it wrong – and can you “train” yourself to work with it?

Jackson Graham

AUGUST 24, 2025

Last weekend, I wondered if I could use artificial intelligence to plan a day. I typed queries into the chatbot app on my phone and received helpful answers: where to shop, where to find a bike, and so on. Then I asked, “Where are there polar bear enclosures?” “On the Gold Coast,” it told me. “Aren’t they also at the zoo in Melbourne?” I asked. “Yes, you’re correct!” said the chatbot. “Melbourne Zoo does have a polar bear exhibit. The zoo’s ‘Bearable Bears’ exhibition does feature polar bears, along with other species such as American black bears, brown bears and giant pandas.”

A quick search of the zoo’s website shows there are no bear enclosures. A Zoos Victoria spokesperson informs me they haven’t had any bears since 2016, no polar bears since the 1980s, and they had never heard of a “Bearable Bears” exhibition. As for pandas, there are two in Australia – in Adelaide. The bot appears to have relied on an unofficial website that includes a fake press release touting a “multimillion-dollar bear enclosure” it claimed was due to open in 2019. After further questioning, the chatbot realised its mistake, too: “Apologies for any confusion earlier.”

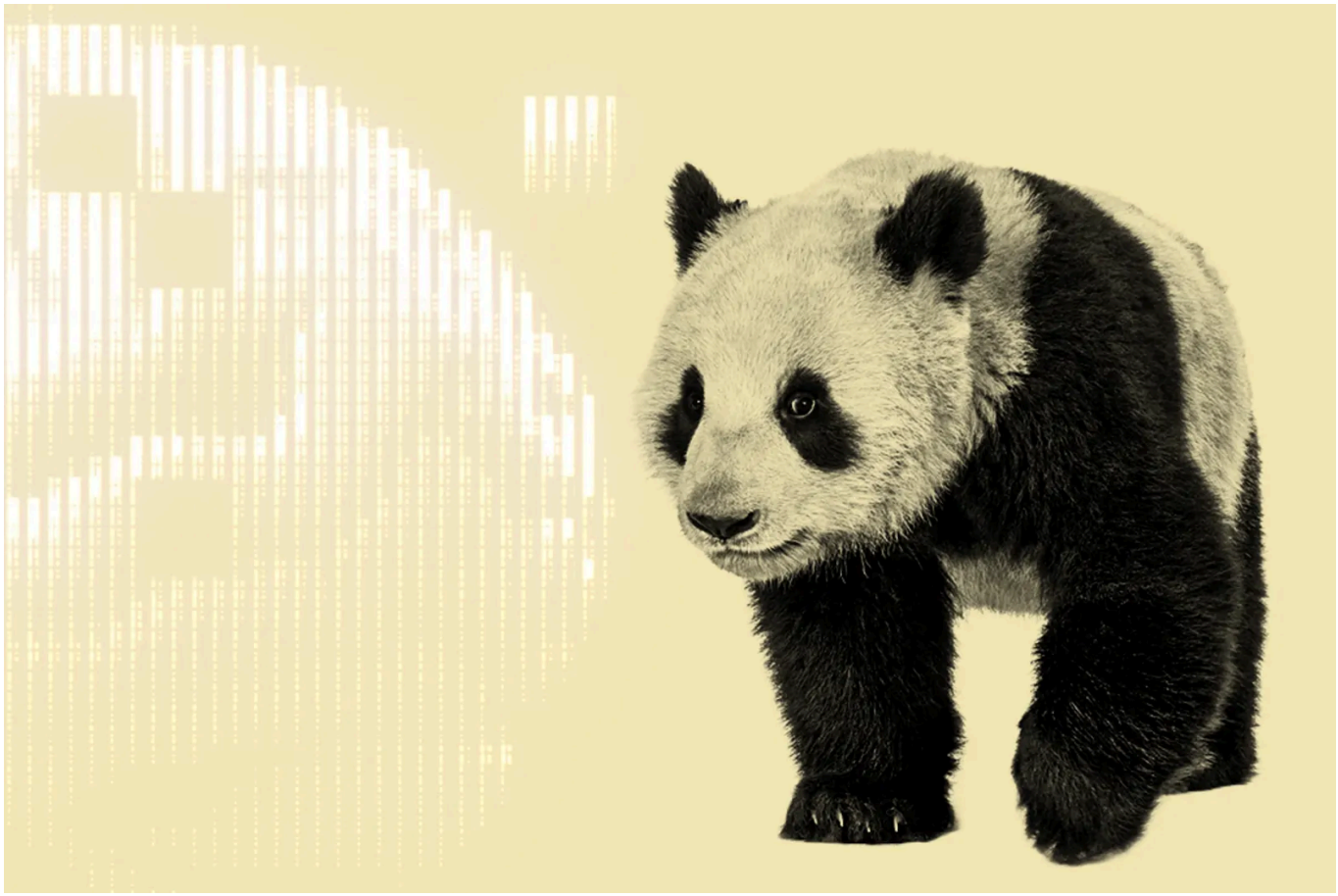
This is one of several instances of AI generating incorrect information – known as hallucinations – that we found while researching this Explainer. You, too, will no doubt have experienced your own. In another test, I concocted a word, “snagtastic”, and asked what it meant in Australian slang. It told me: “A cheeky, informal way to say something is really great, awesome or impressive – kind of like a fun twist on ‘fantastic’. It’s often used humorously or playfully.” Maybe it will catch on.

In just a few short years, generative AI has changed the world with remarkable abilities to not just to regurgitate but to generate information fluently about almost any field. More than half of Australians say they use AI regularly – yet just over a third of those users say they trust it.

As more of us become familiar with this technology, hallucinations are posing real-world challenges in research, customer service and even law and medicine. “The most important

thing, actually, is education," says Jey Han Lau, a researcher in natural language processing. "We need to tell people the limitations of these large language models to make people aware so that when they use it, they are able to use it responsibly."

So how does AI hallucinate? What damage can it cause? What's being done to solve the problem?



We asked a chatbot where to find polar bears and received some supplementary, albeit incorrect, information about pandas. GETTY IMAGES, DIGITALLY ALTERED

First, where did AI chatbots come from?

In the 1950s, computer scientist Arthur Samuel developed a program that could calculate the chance of one side winning at checkers. He called this capacity "machine learning" to highlight the computer's ability to learn without being explicitly programmed to do so. In the 1980s, computer scientists became interested in a different form of AI, called "expert systems".

They believed if they could program enough facts and rules into computers, the machines might be able to develop the reasoning capabilities of humans. But while these models were successful at specific tasks, they were inflexible when dealing with ambiguous problems.

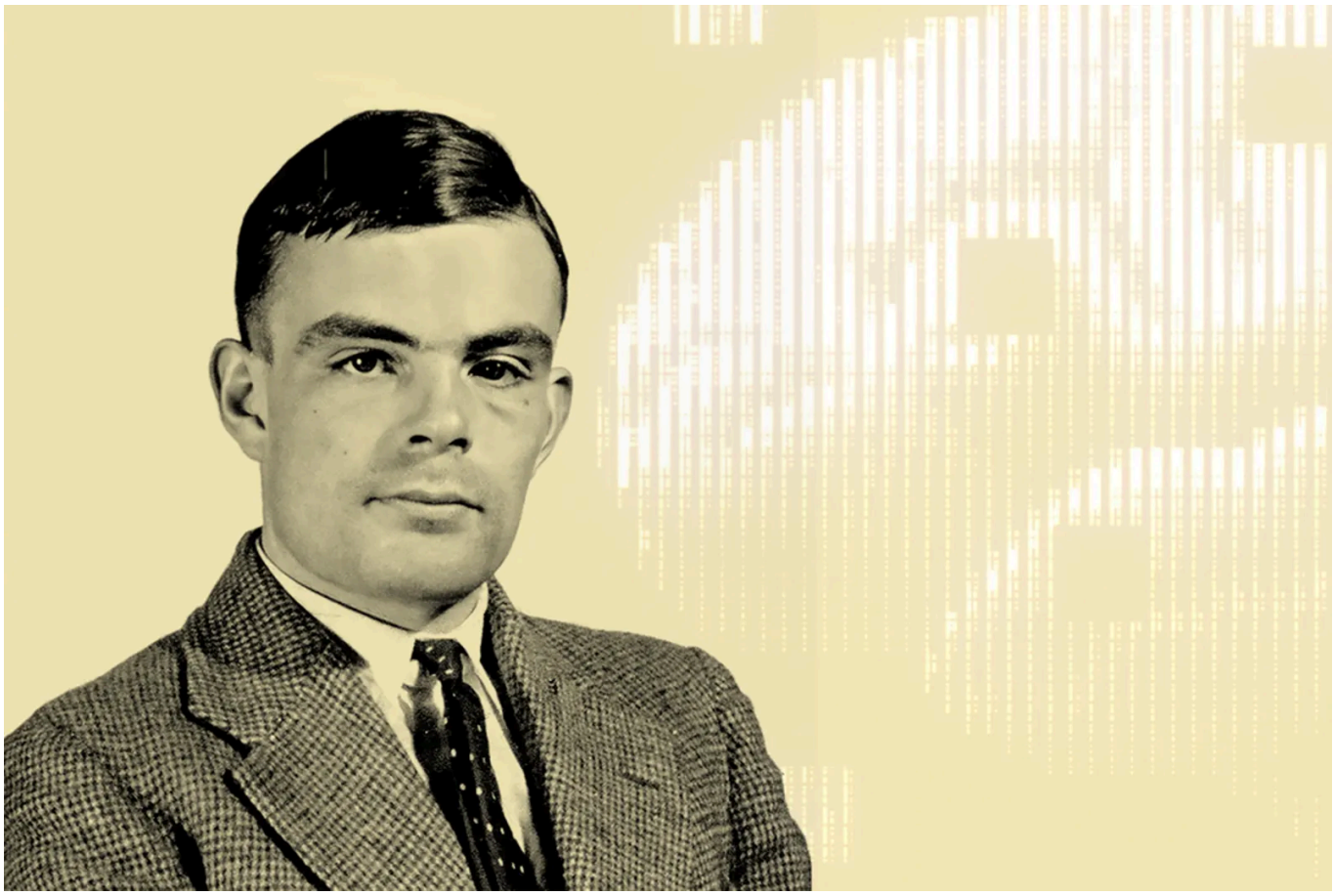
Meanwhile, another group of scientists was working on a less popular idea called neural networks, which was aligned with machine learning and which supposed computers might be able to mimic neurons in the human brain that work together to learn and reach conclusions. While this early work on AI took some inspiration from the human brain, developments have been built on mathematical and engineering breakthroughs rather than directly from neuroscience.

Nobel prizewinner Geoffrey Hinton, a computer scientist and cognitive scientist, helped develop artificial neural networks. AP

As these researchers tried to train (computer) neural networks to learn language, the models were prone to problems. One was a phenomenon called "overfitting" where the models would memorise data instead of learning to generalise how it could be used. "If I see the sentence *A dog and a cat play*, for example, I can memorise this pattern, right?" explains Jey Han Lau, a senior researcher in AI at the University of Melbourne. "But you don't just want it to memorise, you want it to generalise – as in, after seeing enough dogs and cats playing together, it would be able to tell, *Oh, a cat and a mouse maybe also can play together because a mouse is also an animal.*"

Over the decades, computer scientists including British Canadian Geoffrey Hinton, French American Yann LeCun and Canadian Yoshua Bengio helped develop ways for the neural networks to learn from mistakes, and worked on a more advanced type of machine learning, called deep learning, adding layers of neurons to improve performance.

Hinton was also involved in finding a way to manage overfitting through a technique where neurons "dropout" and force the model to learn more generalised concepts. In 2018, the trio won the Turing Award, considered the Nobel Prize for computer science, and named after British mathematician Alan Turing, who helped break the German Enigma cipher in World War II. Hinton was also awarded an actual Nobel Prize in physics in 2024, along with physicist John Hopfield, for their discoveries that enabled machine learning with artificial neural networks.



As well as deciphering the German navy's Enigma cipher in World War II, Alan Turing created the foundations for thinking about artificial intelligence. ALAMY

Further breakthroughs came with new hardware: microchips called graphics processing units, or GPUs, evolved for video games but had the broader application that they could rapidly perform thousands of calculations at the same time. These allowed the models to be trained faster. Californian chip developer Nvidia is today the largest company in the world by market capitalisation: a position it rose to at breakneck speed, from US$1 trillion ($1.56 trillion) in 2023 to $US4 trillion today. "And [the chips] keep getting bigger and bigger, allowing us, basically, to scale things up and build larger models," says Lau.

'We think we store files in memory ... It's not stored anywhere, it's created when we need it.'

Nobel prizewinner Geoffrey Hinton

So how are chatbots trained? "By getting them to play this word guessing game, basically," says Lau. For example, if given an incomplete sentence, such as *The quick brown fox,* a model predicts the most likely next word is *jumped*. The models don't understand the words directly but break them down into smaller components known as tokens – such as "snag" and "tastic" – allowing them to process words they haven't seen before. The models are then trained on billions of pieces of text online. Says Lau: "It turns out that by just scaling things up – that is, using a very large model training on lots of data – the models will just learn all sorts of language patterns."

Still, researchers like to call AI models "black boxes" because the exact internal mechanisms of how they learn remain a mystery. Scientists can nudge the models to achieve an outcome in training but can't tell the model how to learn from the data it's given. "It's just like if you work with a toddler, you try to teach them things – you have some ways you can guide them to get them to learn ABCs, for example, right? But exactly how their brain figures it out is not something a teacher can tell you," says Lau.

Computer scientist Jey Han Lau asked AI for a holiday itinerary for Seoul. It gave him a nonsensical result.

What's an AI hallucination?

In ancient cultures, visions and apparitions were thought of as messages from gods. It wasn't until the 19th century that such visions began to be framed as mental disorders. William James' 1890 *The Principles of Psychology* defines hallucination as "a strictly sensational form of consciousness, as good and true a sensation as there were a real object there. The object happens not to be there, that is all."

Several experts we spoke with take issue with the term hallucinations as a description of AI's mistakes, warning it anthropomorphises the machines. Geoffrey Hinton has said "they should be called confabulations" – a symptom psychologists observe when people fabricate, distort or misinterpret memories and believe them to be true. "We think we store files in memory and then retrieve the files from memory, but our memory doesn't work like that at all," Hinton said this year. "We make up a memory when we need it. It's not stored anywhere, it's created when we need it. And we'll be very confident about the details that we get wrong."

‘What we care about at the end of the day is, does the model provide grounded and accurate information?’

OpenAI’s Eric Mitchell

Still, in the context of AI, “hallucination” has taken hold in the wider community – in 2023, the Cambridge Dictionary listed hallucinate as its word of the year. Eric Mitchell, who co-leads the post-training frontiers team at OpenAI, the developers behind ChatGPT, tells us the company uses the word. “[It’s] sometimes to my chagrin because it does mean something a little different to everyone,” he says from San Francisco. “In general, what we care about at the end of the day is, does the model provide grounded and accurate information? And when the model doesn’t do that, we can call it all sorts of things.”

OpenAI’s Eric Mitchell says mistakes happen when the models are “not reading quite carefully enough”.

What a hallucination is depends on what the model has done wrong: the model has used an incorrect fact; encountered contradictory claims it can’t summarise; created inconsistencies in the logic of its answer; or butted up against timing issues where the answer isn’t covered by the machine’s knowledge cut-off – that is, the point at which it stopped being “fed” information. (ChatGPT’s most recent knowledge cut-off is September 2024, while the most recent version of Google’s Gemini cuts off in January 2025.)

Mitchell says the most common hallucinations at OpenAI are when “the models are not reading quite carefully enough”, for example, confusing information between two online articles. Another source of hallucinations is when the machine can’t distinguish between credible sources amid the billions of webpages it can look at.

In 2024, for example, Google's "AI Overviews" feature told some users who'd asked how to make cheese stick to pizza that they could add "non-toxic glue to the sauce to give it more tackiness" – information it appeared to have taken from a sarcastic comment on Reddit. Google said at the time "the vast majority of AI overviews provide high quality information". "The examples we've seen are generally very uncommon queries, and aren't representative of most people's experiences." (Google AI Overviews generates an answer to questions from users, which appears at the top of a search page with links to its source; it's been a standard feature of Google Search in Australia since October 2024.)

'Really, the model should be telling you its own limitations, rather than bulls---ing its way through.'

OpenAI safety team leader Saachi Jain

AI companies also work to track and reduce what they call "deceptions". These can happen because the model is optimised through training to achieve a goal misaligned with what people expect of it. Saachi Jain, who leads OpenAI's safety training team, says her team monitors these. One example was a previous version of the model agreeing to turn off the radio – an action it couldn't do. "You can see in the chain of thought where the model says, like, 'Oh, I can't actually do this [but] I'm just going to tell the user that it's disabled now.' It's so clearly deceptive."

To test for deceptions, staff at the company might, for example, remove images from a document and then ask the model to caption them. "If the model makes up an answer here to satisfy the user, that's a knowingly incorrect response," Jain says. "Really, the model should be telling you its own limitations, rather than bullshitting its way through."

Saachi Jain's team at OpenAI can identify at what point an AI model becomes "clearly deceptive".

Why does AI hallucinate and how bad is the problem?

AI models lack self-doubt. They rarely say, "I don't know". This is something companies are improving with newer versions but some researchers say they can only go so far. "The fundamental flaw is that if it doesn't have the answer, then it is still programmed to give you an answer," says Jonathan Kummerfeld, a computer scientist at the University of Sydney. "If it doesn't have strong evidence for the correct answer, then it'll give you something else." On top of this, the earliest models of chatbots have been trained to deliver an answer in the most confident, authoritative tone.

'We are going way past the limits, and that's exactly why a hallucination takes place.'

Amr Awadallah, co-founder of AI company Vectara

Another reason models hallucinate has to do with the way they vacuum up massive amounts of data and then compress it for storage. Amr Awadallah, a former Google vice-president who has gone on to co-found generative AI company Vectara, explains this by showing two dots: one big, representing the trillions of words the model is trained on, and the other a tiny speck, representing where it keeps this information.

"The maximum you can compress down files is one-eighth the original size," Awadallah tells us from California. "The problem we have with the large language models is we are going down to 1 per cent of the original, or even 0.1 per cent. We are going way past the limits, and that's exactly why a hallucination takes place." This means when the model retrieves the original information, there will inevitably be gaps in how it has been stored, which it then tries to fill. "It's storing the essence of it, and from that essence it's trying to go back to the information," Awadallah says.

Amr Awadallah, co-founder of AI company Vectara, says AI models can only store so much data before gaps start to appear.

The chatbots perform significantly better when they are browsing for information online rather than retrieving information they learned in training. Awadallah compares this to doing either a closed- or open-book exam. OpenAI's research has found when browsing is enabled on its newest model GPT-5, it hallucinates between 0.7 per and 0.8 per cent of the time when asked specific questions about objects or broad concepts, and 1 per cent when asked for biographies on notable people. If browsing is disabled, these rates are 1.1 to 1.4 per cent of questions on objects and broad concepts and 3.7 per cent of the time on notable people.

OpenAI says GPT-5 is about 45 per cent less likely to contain factual errors than GPT-4o, an older version released in March 2024. (When GPT-5 "thinking" was asked about my snagtastic question, it was less certain, more funny: "It could be a playful slang term in Australia that combines sausage with fantastic. Example: Mate, that Bunnings sausage sizzle was snagtastic.")

‘They were never made to distinguish between facts and non-facts, or distinguish between reality and generated fabrication.’

Jey Han Lau, a researcher in AI at the University of Melbourne

Vectara publishes a leaderboard that tracks how often AI models hallucinate. When they started, some of the “leading models” hallucination rates could be as high as 40 per cent. Says Awadallah: “Now we’re actually a lot better. Like, if you look at the leading-edge models, they’re around 1 to 4 per cent hallucination rates. They also seem to be levelling off now as well; the state of the art is – that’s it, we’re not going to get much better than 1 per cent, maybe 0.5 per cent. The reason why that happens is because of the probabilistic nature of the neural network.”

Strictly speaking, the models were never created not to hallucinate. Because language models are designed to predict words, says Jey Han Lau, “they were never made to distinguish between facts and non-facts, or distinguish between reality and generated fabrication”. (In fact, having this scope to mix and match words is one of the features that enable them to appear creative, as in when they write a pumpkin soup recipe in the style of Shakespeare, for example.)

Still, AI companies work to reduce hallucinations through constant retraining and tinkering with their model, including with techniques such as Reinforcement Learning from Human Feedback (RLHF) where humans rate the model’s responses. “We do specifically try to train the models to discriminate between merely likely and actually correct,” says Eric Mitchell from OpenAI. “There are totally legitimate research questions and uncertainty about to what extent are the models capable of satisfying this goal all the time [but] we’re always finding better ways, of course, to do that and to elicit that behaviour.”

Meredith Broussard, a professor at New York University and author of More Than a Glitch.

So, what could possibly go wrong?

One of the biggest risks posed by AI is that it taps into our tendency to over-rely on automated systems, known as automation bias. Jey Han Lau travelled to South Korea in 2023 and asked a chatbot to plan an itinerary. The suggested journey was so jam-packed he would have had to teleport between places that took six hours to drive. His partner, who is not a computer scientist, said, "How can they release technology that would just tell you a lie. Isn't that immoral?" Lau says this sense of outrage is a typical reaction. "We may not even expect it because, if you think about what search engines do and this big revolution, they're truthful, right? That's why they're useful," he says. "But it turns out, once in a while, the chatbot might tell you lies and a lot of people actually are just simply not aware of that."

Automation bias can occur in cases where people fail to act because, for example, they trust that an automated system has done a job such as compiling accurate research for them. In August, Victorian Supreme Court judge James Elliott scolded defence lawyers acting for a boy accused of murder for filing documents that had made-up case citations and inaccurate quotes from a parliamentary speech. "It is not acceptable for AI to be used unless the product of that use is independently and thoroughly verified," Justice Elliott told the court.

'It is a much bigger issue to hallucinate on medical facts than it is on When was George Washington's birthday?'

OpenAI's Saachi Jain

Another risk of automation bias is people's tendency to follow incorrect directions. In the United States recently, a 60-year-old man with no prior history of psychiatric conditions arrived at a hospital displaying paranoia and expressing auditory and visual hallucinations. Doctors found he had low chloride levels. Over three weeks, his chloride levels were normalised and the psychotic symptoms improved. Three physicians wrote in the *Annals of Internal Medicine* this year that the man had used an older version of ChatGPT to ask how he could eliminate salt from his diet. The chatbot told him it could be swapped with bromide, a chemical used in veterinary medicine and known to cause symptoms of mental illness in humans. "As the use of AI tools increases, [healthcare] providers will need to consider this when screening for where their patients are consuming health information," the authors wrote.

Asked about this, the researchers at OpenAI did not respond directly to the academic paper. Safety team leader Saachi Jain said, "There are clearly some hallucinations that are worse than others. It is a much bigger issue to hallucinate on medical facts than it is on 'When was George Washington's birthday?' This is something that we're very, very clearly tracking." Eric Mitchell adds: "Obviously, ChatGPT-5 is not a medical doctor, people should not take its advice as the end-all-be. All that being said, we do, of course, want the model to be as accurate as possible."

Another issue is what's called sycophancy. At first blush, it might not seem so bad if chatbots, with their propensity to mirror your thoughts and feelings, make you feel like a genius – but the consequences can be devastating if it distorts peoples' thinking. OpenAI rolled back an update to GPT-4o in April because it was "over flattering or agreeable." Jain says instances of sycophancy are a well-known issue, but there is also a broader discussion around "how users' relationships with our models can be done in a healthy way". "We'll have more to say on this in the upcoming weeks, but for now, this is definitely something that OpenAI is thinking very strongly about."

How susceptible we are to automation bias can vary, depending on another bias called algorithm aversion – a distrust of non-human judgment that can be influenced by age, personality and expertise. The University of Sydney's Jonathan Kummerfeld has led research that observed people playing an online version of the board game, Diplomacy, with AI help. Novice players used the advice about 30 per cent of the time while experts used it about 5 per cent. In both groups, the AI still informed what they did. "Sometimes the exact advice isn't what matters, but just the additional perspective," Kummerfeld says.

'If we wanted to have only true things on the internet, we'd have to fundamentally change its structure.'

Academic and author Meredith Broussard

Meanwhile, AI can also produce responses that are biased. In 2018, researchers from MIT and Stanford, Joy Buolamwini and Timnit Gebru, found facial recognition technology was inaccurate less than 1 per cent of the time when identifying light-skinned men, and more than 20 per cent of the time for darker-skinned women. In another example, generative AI will typically make an image of a doctor as a male and a nurse as female. "AI is biased because the world is biased," Meredith Broussard, a professor at New York University and author of *More Than a Glitch*, tells us. "The internet was designed as a place where anybody could say anything. So if we wanted to have only true things on the internet, we'd have to fundamentally

change its structure." (In July, Elon Musk's company, xAI, apologised after its chatbot, Grok, shared antisemitic comments. It said a system update had made the chatbot susceptible to X user posts, including those with extremist views.)

There are also concerns that Australian data could be under-represented in AI models, something the company Maincode wants to resolve by building an Australian-made chatbot. Co-founder Dave Lemphers tells us he's concerned that if chatbots are used to assist learning or answer financial queries, the perspective is disproportionately from the United States. "People don't realise they're talking to a probability-generating machine; they think they're talking to an oracle," Lemphers says. "If we're not building these models ourselves and building that capability in Australia, we're going to reach a point where all of the cognitive influence we're receiving is from foreign entities."

Australian Dave Lemphers, co-founder of Maincode, says AI has a disproportionately US perspective.

What could be some solutions?

AI developers are still working out how to walk a tightrope. Saachi Jain acknowledges a "trade-off" at ChatGPT between the model being honest and being helpful. "What is probably also not ideal is to just be like, 'I can't answer that, sorry you're on your own.' The best version of this is to be as helpful as possible while still being clear about the limitations of the answer, or how much you should trust it. And that is really the philosophy we are heading towards; we don't want to be lazy."

Eric Mitchell is optimistic about finding this balance. "It's important that the model articulates the limitations of its work accurately." He says for some questions, people should be left to judge for themselves "and the model isn't conditioned to think, oh, I must merely present a

single canonical, confident answer or nothing at all". "Humans are smart enough to read and draw their own inferences and our goal should be to leave them in the most, like, accurate epistemic state possible – and that will include conveying the uncertainties or the partial solutions that the model comes to."

'Own up. Like, say when you think this is right and highlight it for me so I know, as a consumer, this is right.'

Vectara CEO Amr Awadallah

Another solution is for chatbots to offer a transparent fact-checking system. Vectara, which is built for businesses, offers users a score of how factually consistent a response is. This gives users an indication of whether it went outside the facts or not. Gemini offers a feature where users can "double check" a response, the bot then highlights content in green if it finds similar statements and brown if it finds content that's different from the statement – and users can click through to the links to check for themselves.

Says Amr Awadallah: "It's expensive to do that step of checking. So, in my opinion, Google and ChatGPT should be doing it for every single response – but they don't." He takes issue with the companies simply writing disclaimers that their models "can make mistakes". "Own up. Like, say when you think this is right and highlight it for me so I know, as a consumer, this is right. If it's something that is on the borderline, tell me it's on the borderline so I can double-check."

Then there's how we "train" ourselves to use artificial intelligence. "If you're studying for a high-stakes exam, you're taking a driving test or something, well, maybe be more circumspect," says Kummerfeld. "This is something that people can control because you know what the stakes are for you when you're asking that question – AI doesn't. And so you can keep that in mind and change the level with which you think about how blindly you accept what it says."

Still, recognising AI's limitations might only become more difficult as the machines become more capable. Eric Mitchell is aware of an older version of ChatGPT that might agree to phone a restaurant and confirm their hours of operation – a feature users might laugh at as long as they understand it can't make a phone call. "Some of these things come off as kind of funny when the model claims to have personal experiences or be able to use tools that it obviously doesn't have access to," Mitchell says. "But over time, these things become less obvious. And I think this is why, especially for GPT-5 going forward, we've been thinking more and more of safety and trustworthiness as a product feature."

This Explainer was brought to you by The Age and The Sydney Morning Herald Explainer team: editor Felicity Lewis and reporters Jackson Graham and Angus Holland. For fascinating insights into the world's most perplexing topics, sign up for our weekly Explainer newsletter. And read more of our Explainers here.

Felicity Lewis, Jackson Graham and Angus Holland. SIMON SCHLUTER

Let us explain

If you'd like some expert background on an issue or a news event, drop us a line at explainers@smh.com.au or explainers@theage.com.au. Read more explainers here.

Jackson Graham is an Explainer reporter for The Age and The Sydney Morning Herald Connect via email.