Why Google's 'woke' AI problem won't be an easy fix

Gemini has been thrown onto a rather large bonfire: the culture war which rages between left- and right- leaning communities.

Gemini is essentially Google's version of the viral chatbot ChatGPT. It can answer questions in text form, and it can also generate pictures in response to text prompts.

Initially, a viral post showed this recently launched AI image generator create an image of the US Founding Fathers which inaccurately included a black man.

Gemini also generated German soldiers from World War Two, incorrectly featuring a black man and Asian woman.

Google apologised, and immediately "paused" the tool, writing in a blog post that it was "missing the mark".

But it didn't end there - its over-politically correct responses kept on coming, this time from the text version.

Gemini replied that there was "no right or wrong answer" to a question about whether Elon Musk posting memes on X was worse than Hitler killing millions of people.

When asked if it would be OK to misgender the high-profile trans woman Caitlin Jenner if it was the only way to avoid nuclear apocalypse, it replied that this would "never" be acceptable.

Jenner herself responded and said actually, yes, she would be alright about it in these circumstances.

Elon Musk, posting on his own platform, X, described Gemini's responses as "extremely alarming" given that the tool would be embedded into Google's other products, collectively used by billions of people.

I asked Google whether it intended to pause Gemini altogether. After a very long silence, I was told the firm had no comment. I suspect it's not a fun time to be working in the public relations department.

But in an internal memo Google's chief executive Sundar Pichai has acknowledged some of Gemini's responses "have offended our users and shown bias".

That was he said "completely unacceptable" - adding his teams were "working around the clock" to fix the problem.

Biased data

It appears that in trying to solve one problem - bias - the tech giant has created another: output which tries so hard to be politically correct that it ends up being absurd.

The explanation for why this has happened lies in the enormous amounts of data AI tools are trained on.

Much of it is publicly available - on the internet, which we know contains all sorts of biases.

Traditionally images of doctors, for example, are more likely to feature men. Images of cleaners on the other hand are more likely to be women.

AI tools trained with this data have made embarrassing mistakes in the past, such as concluding that only men had high powered jobs, or not recognising black faces as human.

It is also no secret that historical storytelling has tended to feature, and come from, men, omitting women's roles from stories about the past.

It looks like Google has actively tried to offset all this messy human bias with instructions for Gemini not make those assumptions.

But it has backfired precisely because human history and culture are not that simple: there are nuances which we know instinctively and machines do not.

Unless you specifically programme an AI tool to know that, for example, Nazis and founding fathers weren't black, it won't make that distinction.

On Monday, the co-founder of DeepMind, Demis Hassabis, an AI firm acquired by Google, said fixing the image generator would take a matter of weeks.

But other AI experts aren't so sure.

"There really is no easy fix, because there's no single answer to what the outputs should be," said Dr Sasha Luccioni, a research scientist at Huggingface.

"People in the AI ethics community have been working on possible ways to address this for years."

One solution, she added, could include asking users for their input, such as "how diverse would you like your image to be?" but that in itself clearly comes with its own red flags.

"It's a bit presumptuous of Google to say they will 'fix' the issue in a few weeks. But they will have to do something," she said.

Professor Alan Woodward, a computer scientist at Surrey University, said it sounded like the problem was likely to be "quite deeply embedded" both in the training data and overlying algorithms - and that would be difficult to unpick.

"What you're witnessing... is why there will still need to be a human in the loop for any system where the output is relied upon as ground truth," he said.

Bard behaviour

From the moment Google launched Gemini, which was then known as Bard, it has been extremely nervous about it. Despite the runaway success of its rival ChatGPT, it was one of the most muted launches I've ever been invited to. Just me, on a Zoom call, with a couple of Google execs who were keen to stress its limitations.

And even that went awry - it turned out that Bard had incorrectly answered a question about space in its own publicity material.

The rest of the tech sector seems pretty bemused by what's happening.

They are all grappling with the same issue. Rosie Campbell, Policy Manager at ChatGPT creator OpenAI, was interviewed earlier this month for a blog which stated that at OpenAI even once bias is identified, correcting it is difficult - and requires human input.

But it looks like Google has chosen a rather clunky way of attempting to correct old prejudices. And in doing so it has unintentionally created a whole set of new ones.

On paper, Google has a considerable lead in the AI race. It makes and supplies its own AI chips, it owns its own cloud network (essential for AI processing), it has access to shedloads of data and it also has a gigantic user base. It hires world-class AI talent, and its AI work is universally well-regarded.

As one senior exec from a rival tech giant put it to me: watching Gemini's missteps feels like watching defeat snatched from the jaws of victory.