TechScape: Could AI-generated content be dangerous for our health?

Publication Date: 2024-04-09

Author: Alex Hern

Section: Technology

Tags: Technology, TechScape newsletter, Deepfake, Artificial intelligence (AI), Computing, features

Article URL: https://www.theguardian.com/technology/2024/apr/09/techscape-deepfakes-cognitohazards-science-fiction

Let's talk about sci-fi. Neal Stephenson's 1992 novel Snow Crash is the book that launched a thousand startups. It was the first book to use the Hindu term avatar to describe a virtual representation of a person, it coined the term "metaverse", and was one of Mark Zuckerberg's pieces of required reading for new executives at Facebook a decade before he changed the focus of the entire company to attempt to build Stephenson's fictional world in reality. The plot revolves around an image that, when viewed in the metaverse, hijacks the viewer's brain, maiming or killing them. In the fiction of the world, the image crashes the brain, presenting it with an input that simply cannot be correctly processed. It's a recurring idea in science fiction. Perhaps the first clear example came four years earlier, in British SF writer David Langford's short story BLIT, which imagines a terrorist attack using a "basilisk", images which contain "implicit programs which the human equipment cannot safely run". In a sequel to that story, published in Nature in 1999, Langford draws earlier parallels, even pulling in Monty Python's Flying Circus, "with its famous sketch about the World's Funniest Joke that causes all hearers to laugh themselves to death". The collective fiction project SCP coined the name for such ideas: a cognitohazard. An idea, the very thinking of which can be harmful. And one question that deserves to be taken increasingly seriously is: are cognitohazards real? What you know can hurt you I started thinking about that question this week, as part of our reporting into the efforts to automatically identify deepfakes in a year of elections around the world. Since 2017, when I first heard the term in the context of face-swapped porn, it has been possible to identify AI-generated imagery by examination. But that task has got harder and harder, and now we're on the cusp of it being beyond the ken of even experts in the field. So it's a race against time to build systems that can automatically spot and label such material before it breaks that threshold. But what if labelling isn't enough? From my story: Seeing a watermark doesn't necessarily have the effect one would want, says Henry Parker, head of government affairs at factchecking group Logically. The company uses both manual and automatic methods to vet content, Parker says, but labelling can only go so far. "If you tell somebody they're looking at a deepfake before they even watch it, the social psychology of watching that video is so powerful that they will still reference it as if it was fact. So the only thing you can do is ask how can we reduce the amount of time this content is in circulation?" Could we call such a video a cognitohazard? Something so compellingly realistic that you involuntarily treat it as reality, even if you're told otherwise, seems to fit the bill. Of course, that also describes a lot of fiction. A horror story that sticks with you and leaves you unable to sleep at night, or a viscerally unpleasant scene of graphic violence that makes you feel physically unwell, could be a cognitohazard if the definition is stretched that far. The dominoes fall Perhaps closer to the examples from fiction are techniques that hijack, not our emotions, but our attention. Emotions, after all, are rarely under our control at the best of times; feeling something

you don’t want to feel is almost the definition of a negative emotion. Attention is supposed to be different. It’s something we have conscious control over. We talk of being “distracted” at times, but more serious seizures of attention warrant increasingly medicalised language: “obsession”, “compulsion”, “addiction”. The idea of tech attacking our attention isn’t new, and there’s a whole concept of the “attention economy” underpinning that barrage. In a world of ad-supported media, with businesses increasingly competing not for our money directly but for our time, inherently limited to just 24 hours a day, there’s a huge commercial motivation to grab and keep attention. Some of the tools of the trade that have been developed to achieve that goal certainly feel like they tap into something primal. The glaring red dots of new notifications, the tactility of a pull-to-refresh feed and the constant push of gamification have all been discussed at length. And some, I think, have crossed the line to become real cognitohazards. While they may perhaps only be dangerous to those with a susceptibility for having their attention hijacked, the compulsion feels real. One is a type of game: “clickers” or “idle” games, such as the critically feted Universal Paperclips, condense the reward mechanics of a game down to their simplest structures. So called because they almost literally play themselves, idle games present a dazzling array of timers, countdowns and upgrades, and constantly offer some breakthrough, improvement or efficiency just a couple of seconds away. I have lost whole days of productivity to them, as have many others. Another is a type of content, what I’ve started to think of as “domino videos”, are the non-interactive equivalent of an idle game. A video of some process progressing in an orderly, yet not quite fully predictable, manner, drawing you in and leading to an inexorable compulsion to watch until the end. Sometimes that’s literally a domino run; other times, it might be someone methodically cleaning a carpet or depilling a sweater. Sometimes the process might never even complete; pong wars is an automatically generated “game” of Breakout, with two balls each threatening to invade the others’ space. It never ends, but you will watch it for longer than is worthwhile. There’s a chance that this is as bad as it gets. It may be that there is something inherently off-putting about true attention harvesters, that means that the drive to stare at them as the progress is always going to be counteracted by the shame or disgust at having wasted time. But what if it isn’t? What does it look like if generative AI is set loose on social media to truly capture attention at an industrial scale? If the advice that parents give to young children isn’t just to be careful of who they speak to on the internet, but to be wary of what they even look at? Everything is science fiction until it’s reality. If you want to read the complete version of the newsletter please subscribe to receive TechScape in your inbox every Tuesday. Join Alex Hern for a Guardian Live online event about AI, deepfakes and elections, on Wednesday 24 April at 8pm BST. Book tickets here.