‘You can do both’: experts seek ‘good AI’ while attempting to avoid the bad

Publication Date: 2023-07-07

Author: Hannah Devlin

Section: Technology

Tags: Artificial intelligence (AI), United Nations, Switzerland, Computing, Europe, features

Article URL: https://www.theguardian.com/technology/2023/jul/07/ai-for-good-artificial-intelligence-conference

Humanity is at a crossroads that may be summed up as AI for good v AI gone bad, according to a leading artificial intelligence expert. “I see two futures here,” the author Prof Gary Marcus told the UN’s AI for Good global summit on Friday. In the rosier version, AI revolutionises medicine, helps tackle the climate emergency and delivers compassionate care to elderly people. But we could be on the precipice of a bleaker alternative, with out-of-control cybercrime, devastating conflict and a descent into anarchy. “I’m not saying what’s coming; I’m saying we need to figure out what we’re doing,” Marcus told the summit. During the week-long event, ostensibly focused on the positive, delegates heard wide-ranging examples of harnessing AI for the benefit of humanity. A cast of robot ambassadors, whose roving gazes could feel unnerving face to face, offered new visions for how elderly people could maintain independence for longer or how autistic children could learn about the world without feeling overwhelmed. Google DeepMind’s chief operating officer, Lila Ibrahim, described how the company’s protein folding breakthrough could transform medicine. Werner Vogels, the chief technology officer at Amazon, described a machine vision system for tracking 100,000 salmon kept in a pen together to detect disease. AI-driven fish farming might not be the most heartwarming image, he acknowledged, but could radically reduce the carbon footprint of global food production. In what could be a nod to those who view “AI for good” as mostly a PR exercise, Vogels noted that cutting-edge technologies “have the potential to not only do AI for good, but to do AI for profit at the same time”. Behind the scenes though, roundtable discussions between diplomats and invited delegates focused less on “good AI” and more on the pressing issue of how to avoid the bad. “It’s not enough if Google is doing a bunch of AI for good. They’ve got to also not be evil,” said Prof Joanna Bryson, an ethics and technology expert at the Hertie school in Berlin, who was not attending the conference. “Good and evil might be opposites, but doing good and doing evil are not opposites. You can do both.” This is a risk, some say, even for seemingly positive applications of AI. A robot, tasked with fetching a coffee, say, may plough down everything and everyone in its path to achieve this narrow goal. ChatGPT, although astonishingly adept with language, appears to make things up all the time. “If humans behaved in this way, you’d say they had a kind of psychosis,” said Prof Stuart Russell, an AI pioneer at the University of California, Berkeley. But nobody fully understands the internal workings of ChatGPT and it cannot be readily programmed to tell the truth. “There’s nowhere to put that rule in,” said Russell. “We know how to make AI that people want, but we don’t know how to make AI that people can trust,” said Marcus. The question of how to imbue AI with human values is sometimes referred to as “the alignment problem”, although it is not a neatly defined computational puzzle that can be resolved and implemented in law. This means that the question of how to regulate AI is a massive, open-ended scientific question – on top of significant commercial, social and political interests that need to

be navigated. Scientists and some tech companies are looking at these questions in earnest – but in some cases it is a game of catch-up with technologies that have already been deployed. Marcus used his presentation to launch a Centre for the Advancement of Trustworthy AI, which he hopes will act as a Cern-like, philanthropically funded international agency on the theme. Prof Maja Mataric of the University of Southern California described new research (published on Arxiv) analysing the personalities of large-language models – and how they might be shaped to be prosocial to "keep them safe". "I don't want a weird personality," she said. "Well-designed systems can be good for humanity." Others would like to see a tighter focus on AI that is already in widespread use, rather than far-flung scenarios of superhuman intelligence, which may never materialise. "Mass discrimination, the black box problem, data protection violations, large-scale unemployment and environmental harms – these are the actual existential risks," said Prof Sandra Wachter of the University of Oxford, one of the speakers at the summit. "We need to focus on these issues right now and not get distracted by hypothetical risks. This is a disservice to the people who are already suffering under the impact of AI." Either way there is a growing consensus, among tech companies and governments, that governance is needed – and quickly. "It should be done pretty fast … within half a year, a year," said Dr Reinhard Scholl of the UN's International Telecommunication Union and co-founder of the AI for Good Summit. "People agree that if you have to wait for a few years that would not be good."