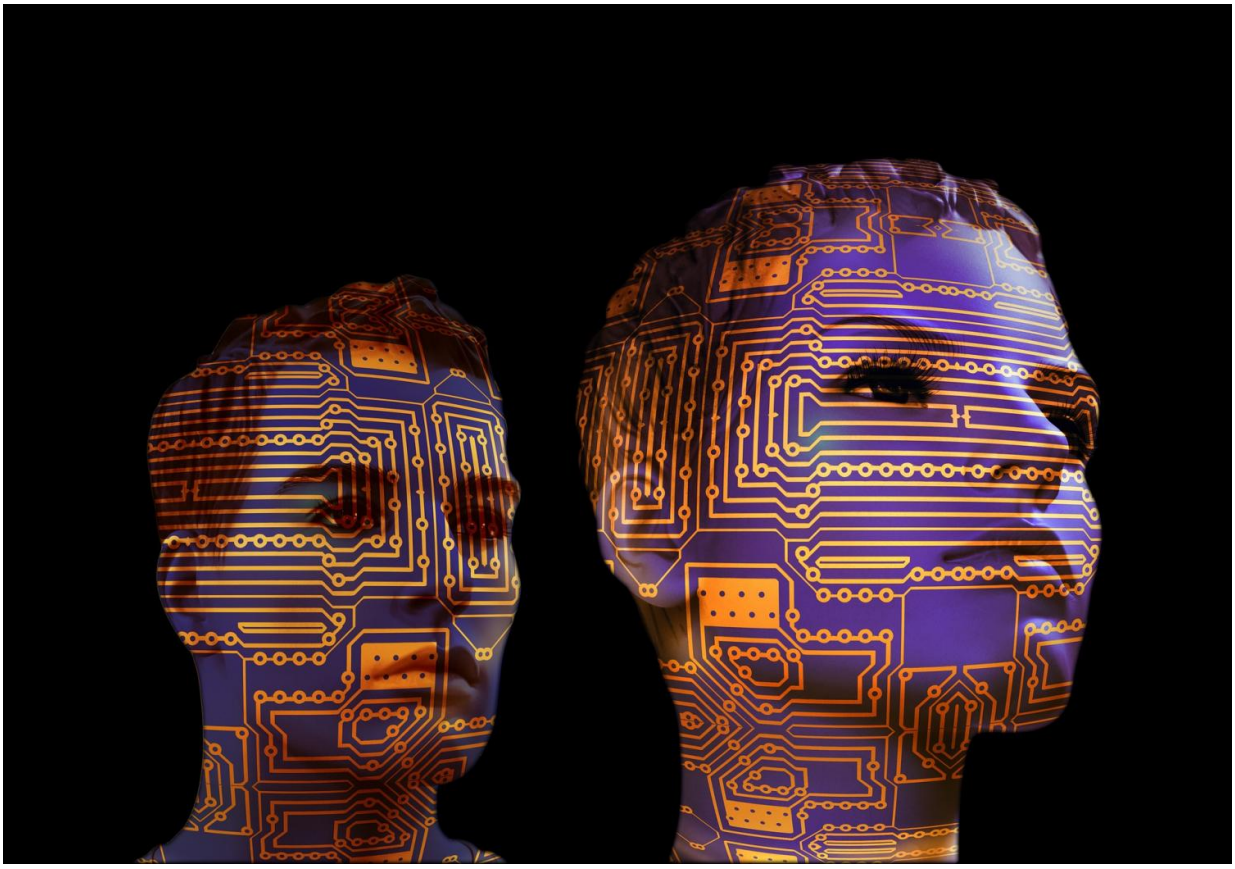
Google collaborates with others over Artificial Intelligence safety

June 23 2016, by Nancy Owano



Credit: CC0 Public Domain

(Tech Xplore)—The Google Research Blog on Tuesday posted a message from Chris Olah of Google Research. He said, " today we're publishing a technical paper, Concrete Problems in AI Safety, a

collaboration among scientists at Google, OpenAI, Stanford and Berkeley."

News about a research paper? What's the big deal? The big deal is a very big deal for those alarmed over what limits may be over-stepped by AI systems in carrying out their actions, and whether we had better anticipate any event where an AI system does not behave according to a predesigned purpose engineered by humans and where the unintended consequences deliver great harm.

Google believes it is time to move to another rung than just fretting. Said Cade Metz in *Wired* on Tuesday: "...that's kind of the point: Because no one has good answers, it's time to start looking for them."

Olah said, "We believe it's essential to ground concerns in real machine learning research, and to start developing practical approaches for engineering AI systems that operate safely and reliably."

So what does Google suggest? Namely, it is putting in place a push for "open, cross-institution work on how to build machine learning systems that work as intended. We're eager to continue our collaborations with other research groups to make positive progress on AI."

The paper is a collaboration between Google (Google Brain) and Stanford, University of California at Berkeley and OpenAI. The latter is a non-profit artificial intelligence research company.

Paul Christiano and Greg Brockman from the latter group blogged on Tuesday: "Advancing AI requires making AI systems smarter, but it also requires preventing accidents—that is, ensuring that AI systems do what people actually want them to do...We think that broad AI safety collaborations will enable everyone to build better machine learning systems."

Google's Olah and team outlined five problems that the team thinks are quite important as AI becomes applied in more general circumstances. "These are all forward thinking, long-term research questions—minor issues today, but important to address for future systems," he said.

Actually, said *MIT Technology Review*, the company is laying out "five unsolved challenges that need to be addressed if smart machines such as domestic robots are to be safe." Tom Simonite wrote about the Olah blog's' use of a cleaning robot to illustrate some the five points.

"One area of concern is in preventing systems from achieving their objectives by cheating. For example, the cleaning robot might discover it can satisfy its programming to clean up stains by hiding them instead of actually removing them," wrote Simonite.

Another problem is how the AI machine can explore a new environment safely. To use the cleaning example, a cleaning robot should be able to experiment with mopping strategies, "but clearly it shouldn't try putting a wet mop in an electrical outlet," blogged Olah.

As for anyone who still thinks worrying over AI is silly, Cade Metz in *Wired* delivered some thoughts about AI and the future on Wednesday. Metz quoted Alexander Reben, a roboticist and artist in Berkeley, California, who said, "We're starting to get into gray areas. We don't always know which inputs yield which outputs."

More information: — research.googleblog.com/2016/0 … on-to-ai-safety.html

— openai.com/blog/concrete-ai-safety-problems/

— Concrete Problems in AI Safety, arXiv:1606.06565 [cs.AI] arxiv.org/abs/1606.06565