DAILY NEWS 30 August 2017

Fatal AI mistakes could be prevented by having human teachers

Not the ideal time for an AI to make a mistake
Bloomberg/Getty

By **Matt Reynolds**

Artificial intelligence needs our help. The best AIs are quickly mastering skills from lip-reading to video games, but only by learning through repeated failure. As robots take on riskier domains, like healthcare and driving, this is no longer an acceptable approach. Fortunately, a new study suggests that with the right human oversight, it might be possible to ditch the failures.

To try to train an AI without it making a mistake, Owain Evans at the University of Oxford and his colleagues started with the simple two-dimensional table tennis video game *Pong*. Normally, a *Pong*-playing agent will let the ball fly past its paddle a few hundred times before realising that isn't a very good way of increasing its score. But in this case, a human would step in to avoid that happening.

Another AI watched as the human intervened in the game. After observing the human for 4.5 hours, it was then able to mimic the human overseer and prevent the *Pong*-playing AI from making any serious errors in the future.

Evans's study suggests that, given the right circumstances, it is possible to train an AI so it learns a task without experiencing a serious failure. The same approach also worked for training an AI to play *Space Invaders* without making any big mistakes.

Learning from mistakes

A little human oversight isn't just useful for AIs playing computer games. Evans says that if more humans had kept a close eye on Facebook's news-recommending algorithms, it might not have showered us with fake news.

Having a human in the loop doesn't always stop AI going wrong, however. When Evans tried the same approach with the game *Road Runner*, the AI overseer wasn't able to block every big mistake the game-playing AI made. More complicated Atari games would require years of human oversight before agents were able to play without making mistakes.

Even a system trained with human oversight is never going to be absolutely safe. It's hard to know how these systems will behave in circumstances that an AI hasn't been trained to handle, says Evans. And even the best AI could be led astray by a sloppy human trainer. "This is only as good as the human," says Evans.

If we are to trust robots in the home and hospitals, then we will need to have some guarantees about their safety, says David Abel at Brown University in Providence, Rhode Island.

More improvements could come if AIs were trained to deliberately make mistakes early in their training, so their learning advances faster.

Reference: arxiv.org/abs/1707.05173

Read more: DeepMind dojo will train AI to beat human StarCraft players; AI learns to play video game from instructions in plain English