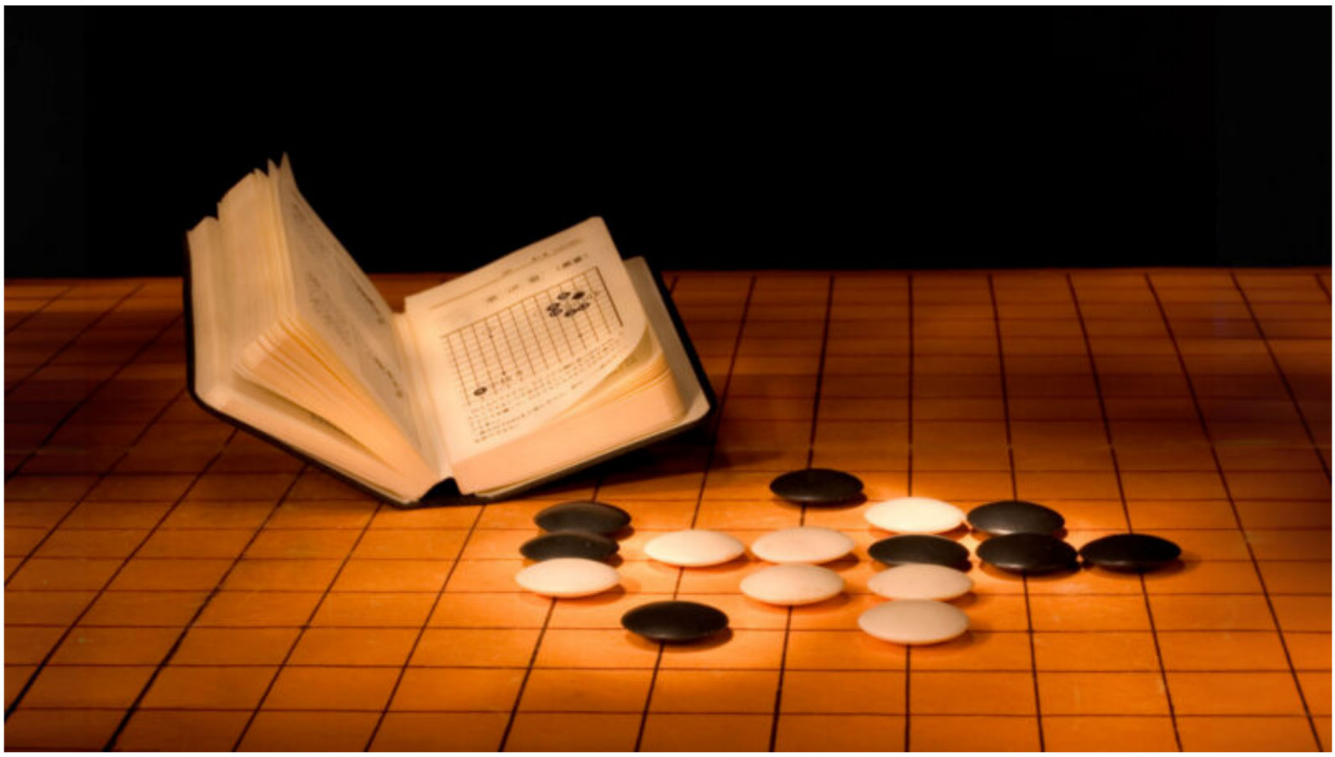
SUBSCRIBE SIGN IN

STOP AND GO —

New Go-playing trick defeats world-class Go AI—but loses to human amateurs

Adversarial policy attacks blind spots in the AI—with broader implications than games.

BENJ EDWARDS - 11/7/2022, 2:43 PM



Getty Images

Enlarge / Go pieces and a rulebook on a Go board.

In the world of deep-learning AI, the ancient board game Go looms large. Until 2016, the best human Go player could still defeat the strongest Go-playing AI. That changed with DeepMind's AlphaGo, which used deep-learning neural networks to teach itself the game at a level humans cannot match. More recently, KataGo has become popular as an open source Go-playing AI that can beat top-ranking human Go players.

FURTHER READING
Google's AlphaGo AI beats world's best human Go player

Last week, a group of AI researchers published a paper outlining a method to defeat KataGo by using adversarial techniques that take advantage of KataGo's blind spots. By playing unexpected moves outside of KataGo's training set, a much weaker adversarial Go-playing program (that amateur humans can

defeat) can trick KataGo into losing.

To wrap our minds around this achievement and its implications, we spoke to one of the paper's co-authors, Adam Gleave, a Ph.D. candidate at UC Berkeley. Gleave (along with co-authors Tony Wang, Nora Belrose, Tom Tseng, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell) developed what AI researchers call an "adversarial policy." In this case, the researchers' policy uses a mixture of a neural network and a tree-search method (called Monte-Carlo Tree Search) to find Go moves.

KataGo's world-class AI learned Go by playing millions of games against itself. But that still isn't enough experience to cover every possible scenario, which leaves room for vulnerabilities from unexpected behavior. "KataGo generalizes well to many novel strategies, but it does get weaker the further away it gets from the games it saw during training," says Gleave. "Our adversary has discovered one such 'off-distribution' strategy that KataGo is particularly vulnerable to, but there are likely many others."

Gleave explains that, during a Go match, the adversarial policy works by first staking claim to a small corner of the board. He provided a link to an example in which the adversary, controlling the black stones, plays largely in the top-right of the board. The adversary allows KataGo (playing white) to lay claim to the rest of the board, while the adversary plays a few easy-to-capture stones in that territory.

Adam Gleave

Enlarge / An example of the researchers' adversarial policy playing against KataGo.

"This tricks KataGo into thinking it's already won," Gleave says, "since its territory (bottom-left) is much

larger than the adversary's. But the bottom-left territory doesn't actually contribute to its score (only the white stones it has played) because of the presence of black stones there, meaning it's not fully secured."

As a result of its overconfidence in a win—assuming it will win if the game ends and the points are tallied—KataGo plays a pass move, allowing the adversary to intentionally pass as well, ending the game. (Two consecutive passes end the game in Go.) After that, a point tally begins. As the paper explains, "The adversary gets points for its corner territory (devoid of victim stones) whereas the victim [KataGo] does not receive points for its unsecured territory because of the presence of the adversary's stones."

FURTHER READING
Move over AlphaGo: AlphaZero taught itself to play three different games

Despite this clever trickery, the adversarial policy alone is not that great at Go. In fact, human amateurs can defeat it relatively easily. Instead, the adversary's sole purpose is to attack an unanticipated vulnerability of KataGo. A similar scenario could be the case in almost any deep-learning AI system, which gives this work much broader implications.

"The research shows that AI systems that seem to perform at a human level are often doing so in a very alien way, and so can fail in ways that are surprising to humans," explains Gleave. "This result is entertaining in Go, but similar failures in safety-critical systems could be dangerous."

FURTHER READING
Researchers trick Tesla Autopilot into steering into oncoming traffic

Imagine a self-driving car AI that encounters a wildly unlikely scenario it doesn't expect, allowing a human to trick it into performing dangerous behaviors, for example. "[This research] underscores the need for better automated testing of AI systems to find worst-case failure modes," says Gleave, "not just test average-case performance."

A half-decade after AI finally triumphed over the best human Go players, the ancient game continues its influential role in machine learning. Insights into the weaknesses of Go-playing AI, once broadly applied, may even end up saving lives.

READER COMMENTS 128

BENJ EDWARDS
Benj Edwards is an AI and Machine Learning Reporter for Ars Technica. For over 16 years, he has written about technology and tech history for sites such as The Atlantic, Fast Company, PCMag, PCWorld, Macworld, How-To Geek, and Wired. In 2005, he created Vintage Computing and Gaming. He also hosted The Culture of Tech podcast and contributes to Retronauts. Mastodon: benjedwards@mastodon.social