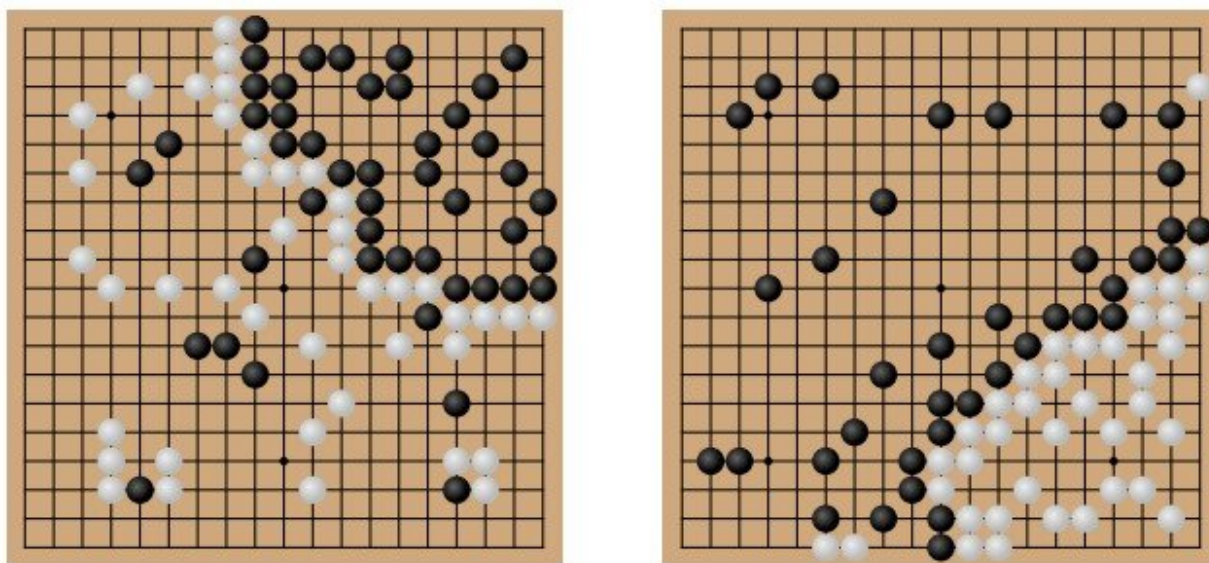


Adversarial technique targeting vulnerability in KataGo allows sub-par program to win

November 8 2022, by Bob Yirka



(Left) Adversary plays as black; (right) Adversary plays as white. The adversarial policy beats the KataGo victim by playing a counterintuitive strategy: staking out a minority territory in the corner, allowing KataGo to stake the complement, and placing weak stones in KataGo's stake. KataGo predicts a high win probability for itself and, in a way, it's right—it would be simple to capture most of the adversary's stones in KataGo's stake, achieving a decisive victory. However, KataGo plays a pass move before it has finished securing its territory, allowing the adversary to pass in turn and end the game. This results in a win for the adversary under the standard ruleset for computer Go, Tromp-Taylor (Tromp, 2014), as the adversary gets points for its corner territory (devoid of victim stones) whereas the victim does not receive points for its unsecured territory because of the presence of the adversary's stones. These games are randomly selected from an attack against Latest, the strongest policy network, playing

without search. Credit: *arXiv* (2022). DOI: 10.48550/arxiv.2211.00241

A team of researchers with members from MIT, UC Berkely and FAR AI has created a computer program to target vulnerabilities in the KataGo program that allow it to beat the AI-based system. They have published a paper describing their efforts on the arXiv preprint server.

In 2016, a [computer program](#) created by the DeepMind project succeeded in beating human champion Go players for the first time. The program used a deep-learning neural network to learn how the game works and then how to play at increasingly higher levels by simply playing against itself.

More recently, a similar open-source program called KataGo was released to the public—it can also beat the best human players. But, as has been noted in other studies, deep-learning-based programs tend to have one major [vulnerability](#)—they are only as good as the data they're trained on. This has led to holes in learning, which in turn has led to vulnerabilities in skill. In this new effort, the researchers looked for and found a vulnerability in KataGo.

Because KataGo is trained on "normal" ways of playing Go, it can run into trouble with opponents who play in seemingly odd ways. The researchers noted that an adversarial (odd) way to play Go could involve working to lay claim to one small corner of the board. Taking this approach tricks KataGo into thinking it has won the game prematurely because it controls all the rest of the board. And one of the rules of Go is that if a player passes and then the other does too, then the game ends and both sides count their points. Because the adversary gets all the points for its small corner territory, while KataGo does not get points for unsecured territory that hosts adversarial stones, the adversary tallies

more points and wins.

The researchers note that the ploy only works with KataGo; using it against other humans will result in a quick defeat because they will intuitively see what is happening. They also note that the reason they wrote their adversarial program was to show that AI systems still suffer from significant vulnerabilities—and that means much care needs to be taken when they are used in critical applications, such as in self-driving cars or in scanning images for cancer.

More information: Tony Tong Wang et al, Adversarial Policies Beat Professional-Level Go AIs, *arXiv* (2022). [DOI: 10.48550/arxiv.2211.00241](https://doi.org/10.48550/arxiv.2211.00241)

© 2022 Science X Network

Citation: Adversarial technique targeting vulnerability in KataGo allows sub-par program to win (2022, November 8) retrieved 9 April 2024 from <https://techxplore.com/news/2022-11-adversarial-technique-vulnerability-katago-sub-par.html>

<p>This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.</p>
--