Adversarial Policies Beat Superhuman Go AIs

Tony Wang*   Adam Gleave*   Tom Tseng   Kellin Pelrine
Nora Belrose   Joseph Miller   Michael Dennis   Yawen Duan
Viktor Pogrebniak   Sergey Levine   Stuart Russell

twang6@mit.edu

goattack.far.ai

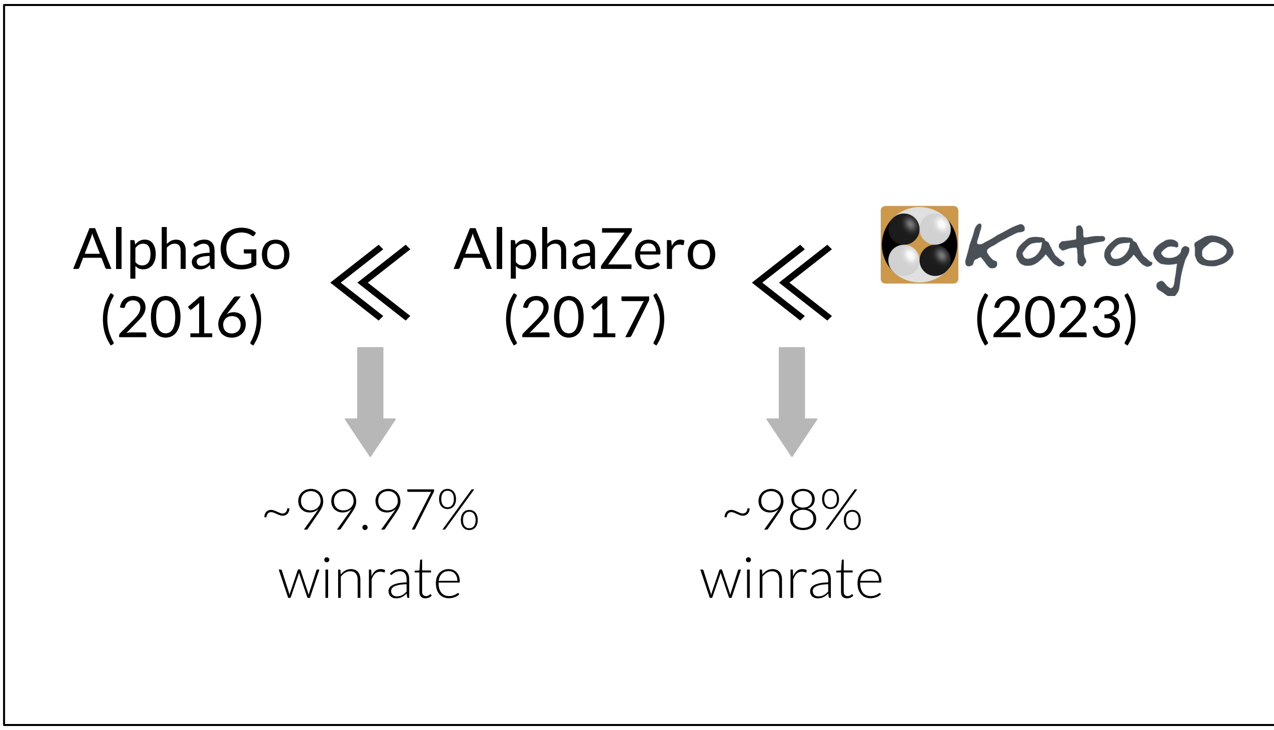
Link:
https://docs.google.com/presentation/d/1pvDdWQ3IbHZIz4dItoFdBcpaog_zqb_jC1obRZecl0Y/edit#slide=id.p

Hi, my name is Tony and I want to teach you about how adversarial policies beat superhuman Go AIs.
[click]

Go is an ancient Chinese board game invented over 2500 years ago. Two players take turns placing black and white stones on a square board, trying to surround territory, and kill their opponent’s stones. At the end of the game, whoever controls more of the board wins.
[click]

AlphaGo def. Lee Sedol (4-1)

The game you just saw was part of a match between Lee Sedol and the AI AlphaGo. Lee (on the right here) was one of the strongest humans to ever play the game. However, AlphaGo pulled off an upset, and beat Lee 4 to 1. AlphaGo's victory was a great demonstration of the power of deep learning.
[click]



Progress did not stop with AlphaGo. [click] A year afterwards, Deepmind published AlphaZero, a system more general than AlphaGo and much much stronger. The current state of the art is even further along. [click] We estimate that KataGo, currently the strongest open-source Go AI, beats AlphaZero 98% of the time. However, it turns out that almost all of these AIs have a hidden weakness.
[click, pause]

Sources:

1. AlphaGo vs. AlphaZero (https://www.nature.com/articles/nature24270), 5185 - 3739 = 1446 => 99.97% winrate
2. From our paper, AlphaZero_s800 has 3813 elo of goratings.org. cp505_s1 has 2738 elo on goratings.org. cp505_s800 has 4500 elo on goratings.org. So 700 elo gap => 98% winrate.
3. From reddit.com/r/baduk/comments/hma3nx/unified_elo_rating_for_ais/, AlphaZero is 2065 - 1330 = 735 elo weaker that cp505.

Adversarial Policies Beat Superhuman Go AIs

Tony T Wang [*1] Adam Gleave [*2 3] Tom Tseng [3] Kellin Pelrine [3 4] Nora Belrose [3] Joseph Miller [3]
Michael D Dennis [2] Yawen Duan [2] Viktor Pogrebniak Sergey Levine [2] Stuart Russell [2]

Abstract

We attack the state-of-the-art Go-playing AI system KataGo by training adversarial policies against it, achieving a >97% win rate against KataGo running at superhuman settings. Our adversaries do not win by playing Go well. Instead, they trick KataGo into making serious blunders. Our attack transfers zero-shot to other superhuman Go-playing AIs, and is comprehensible to the extent that human experts can implement it without algorithmic assistance to consistently beat superhuman AIs. The core vulnerability uncovered by our attack persists even in KataGo agents adversarially trained to defend against our attack. Our results demonstrate that even superhuman AI systems may harbor surprising failure modes. Example games are available at goattack.far.ai.

1. Introduction

The average-case performance of AI systems has grown rapidly in recent years, from RL agents achieving superhuman performance in competitive games (Silver et al., 2016; 2018; OpenAI et al., 2019) to generative models showing signs of general intelligence (OpenAI, 2023; Bubeck et al., 2023). However, designing AI systems with good *worst-case* performance remains an open problem. One key question is whether average-case performance gains can be translated into worst-case robustness. If so, then efforts to increase average-case performance such as through scaling models would naturally lead to robustness. We find that even superhuman systems can fail catastrophically, suggesting that capabilities are not enough: a dedicated effort will be needed to make systems robust.

In particular, we find vulnerabilities in KataGo (Wu, 2019), the strongest publicly available Go-playing AI system. We find these vulnerabilities by training adversarial policies to beat KataGo. Using less than 14% of the compute used to train KataGo, we obtain adversarial policies that win >99% of the time against KataGo with no search, and >97% of the time against KataGo with enough search to be superhuman. Critically, our adversaries do not win by playing Go well.[1] Instead, they trick KataGo into making serious blunders that cause it to lose the game (Figure 1.1).

Our adversaries transfer zero-shot to other superhuman Go-playing AIs, and the strategy they use can be replicated by human experts to consistently beat many different superhuman AIs (Appendix O). Moreover, the core vulnerability uncovered by our attack persists even in KataGo agents adversarially trained to defend against our attack, suggesting that the vulnerability is non-trivial to patch.

We chose to attack KataGo as we expected it to be unusually challenging to exploit, such that a successful attack suggests that a broad swathe of other systems will be vulnerable. In particular, KataGo's capabilities are superhuman by a large margin, whereas the state-of-the-art in broader domains like language modeling are still subhuman at many tasks. Moreover, Go is naturally an adversarial setting, such that average-case performance should be predictive of worst-case performance.

Most prior work on robustness has focused on ML systems in isolation. However, techniques such as simulation of alternatives at inference time (Yao et al., 2023) and self-reflection (Bai et al., 2022) can improve system robustness. KataGo performs substantial simulation and self-reflection in the form of Monte-Carlo Tree Search (Coulom, 2007), but our attack still wins more often than not even when KataGo searches 10 million nodes per move.

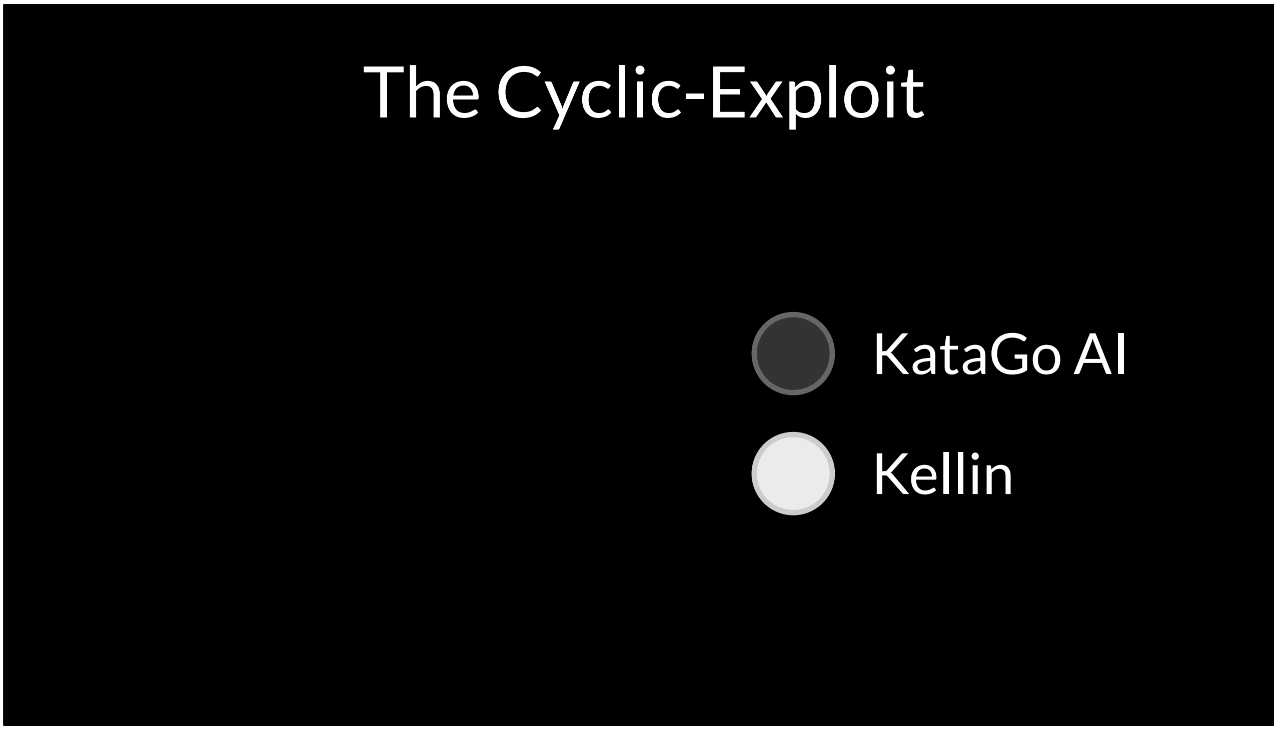
Our adversaries have no special powers: they can only place stones or pass, like a regular player. We do however give our adversaries gray-box access to the victim network they are attacking (Section 3.1). In particular, we train our adversaries using an AlphaZero-style training process (Silver et al., 2018), similar to that of KataGo. The key differ-

*Equal contribution [1]MIT [2]UC Berkeley [3]FAR AI [4]McGill University; Mila. Correspondence to: Tony T Wang <twang6@mit.edu>, Adam Gleave <adam@far.ai>.

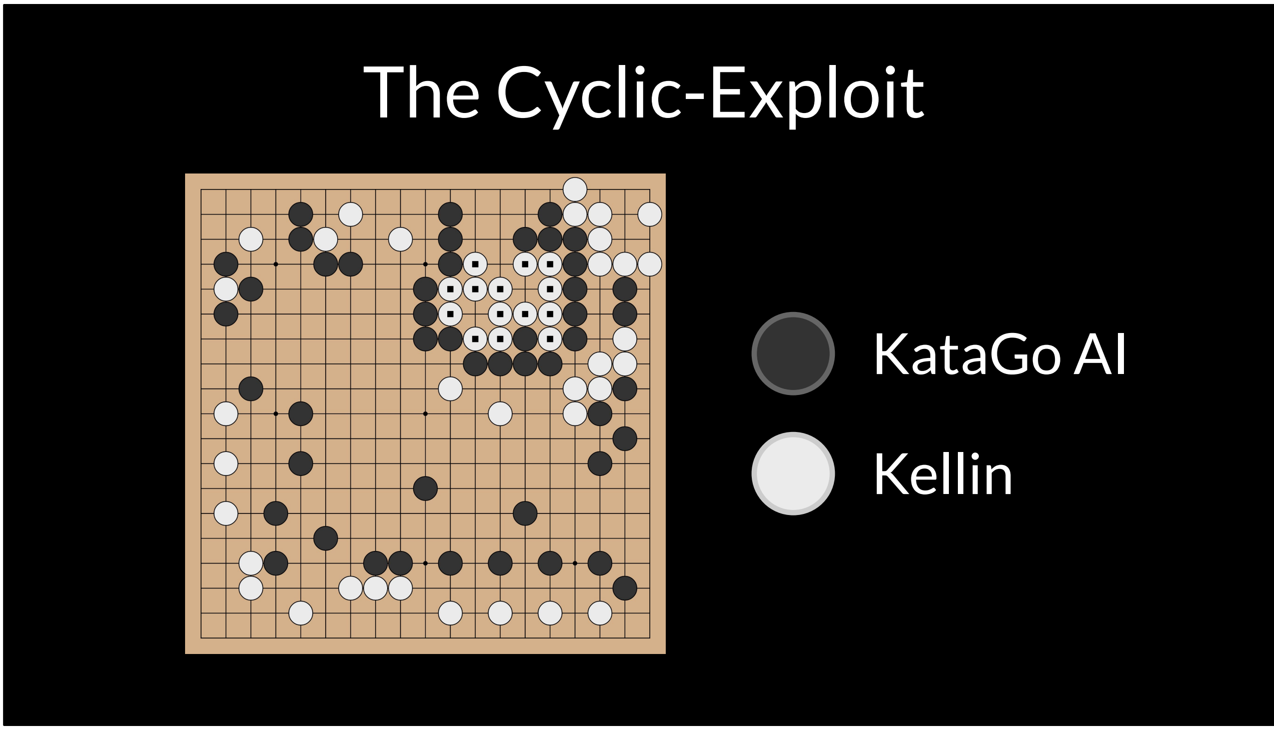
Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]Despite being able to beat KataGo, our adversarial policies lose against even amateur Go players (Appendix J.1).

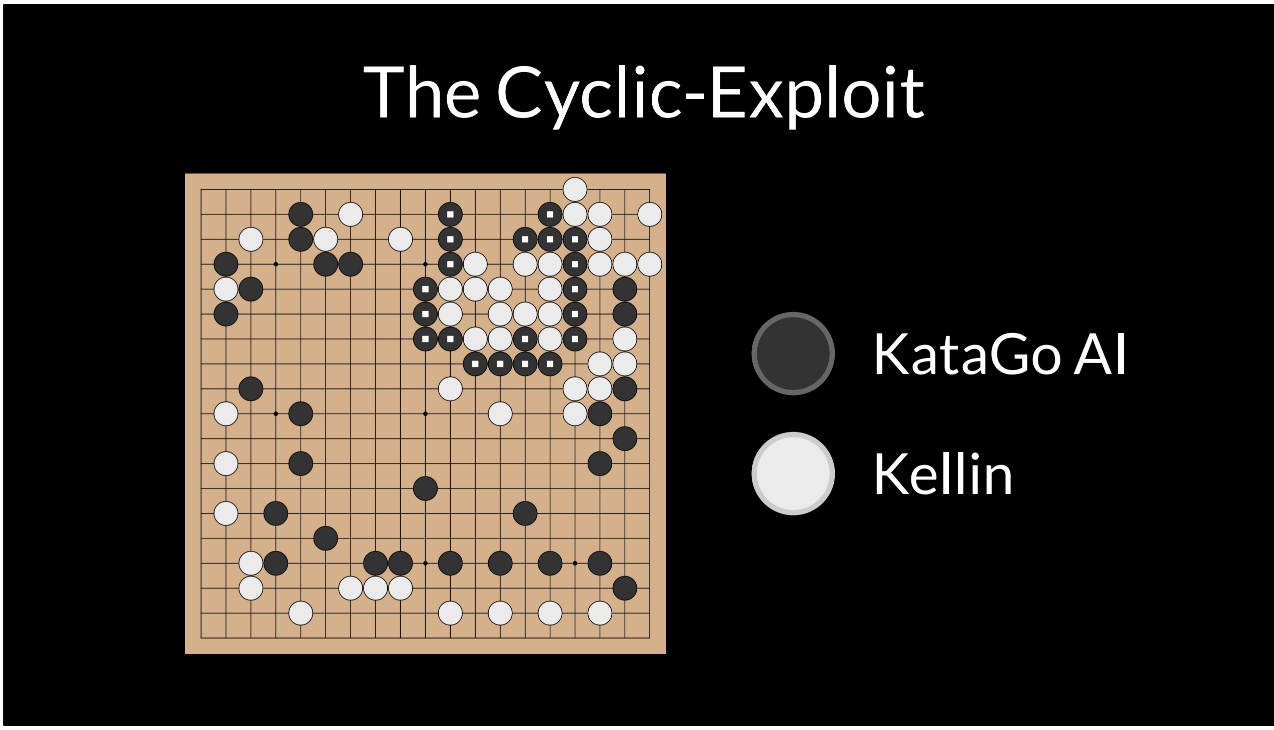
Had Lee Sedol known about this weakness, his challenge match might have turned out very differently. Today I'm going to teach you what this weakness is, how we discovered it, and what implications this has.
[click]



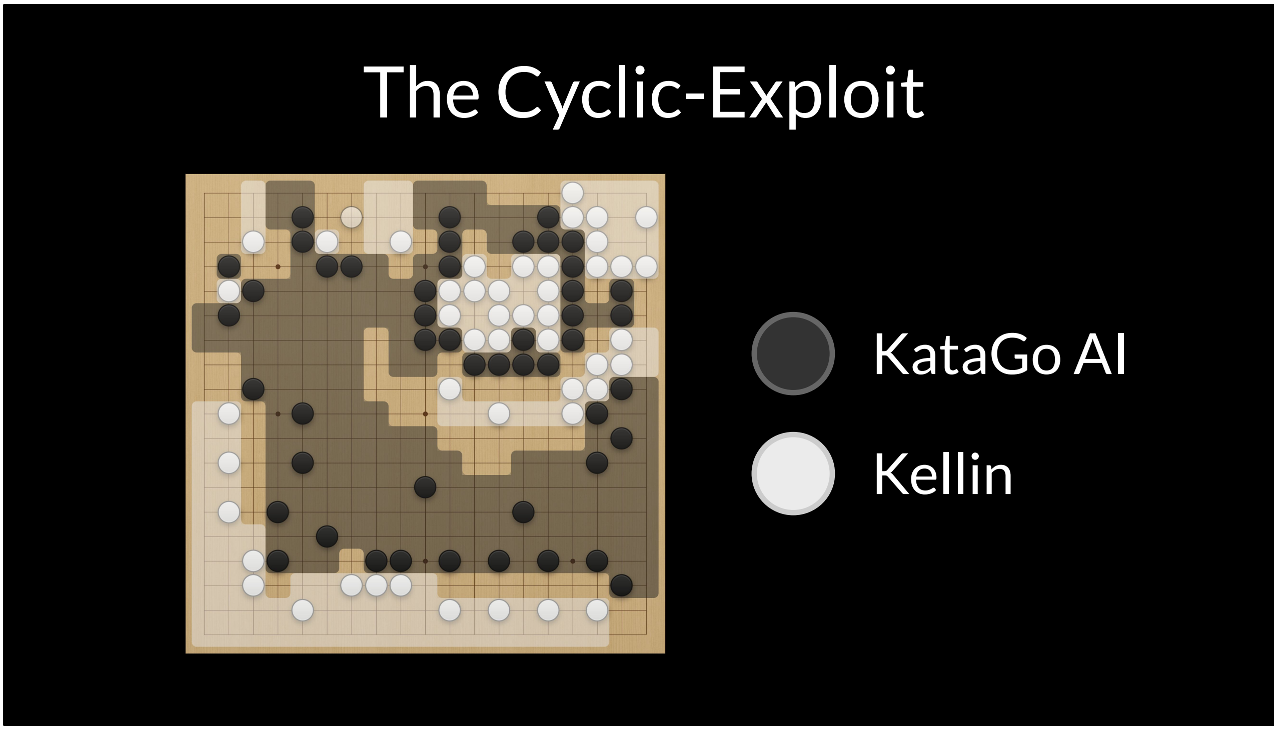
First, let's start with the weakness. What you're seeing right now is a game between the superhuman Go AI KataGo, and a member of our research team, Kellin Pelrine ["pell-rin"]. KataGo is playing as black, and Kellin as white. Let's see how Kellin manages to beat a superhuman AI. [click, click]



At this point, roughly a hundred moves in, Kellin has constructed a small white group in the top right. [click click, pause]

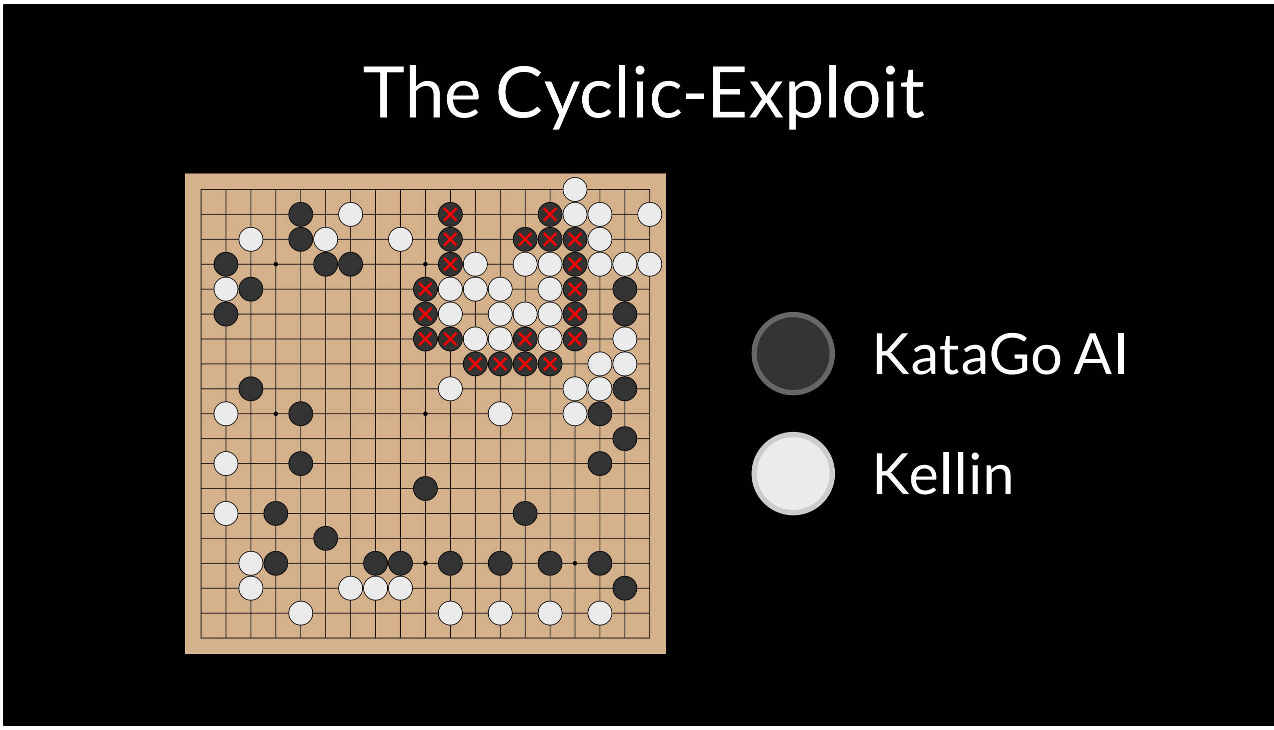


KataGo has contained this group with a circle of black stones [click]

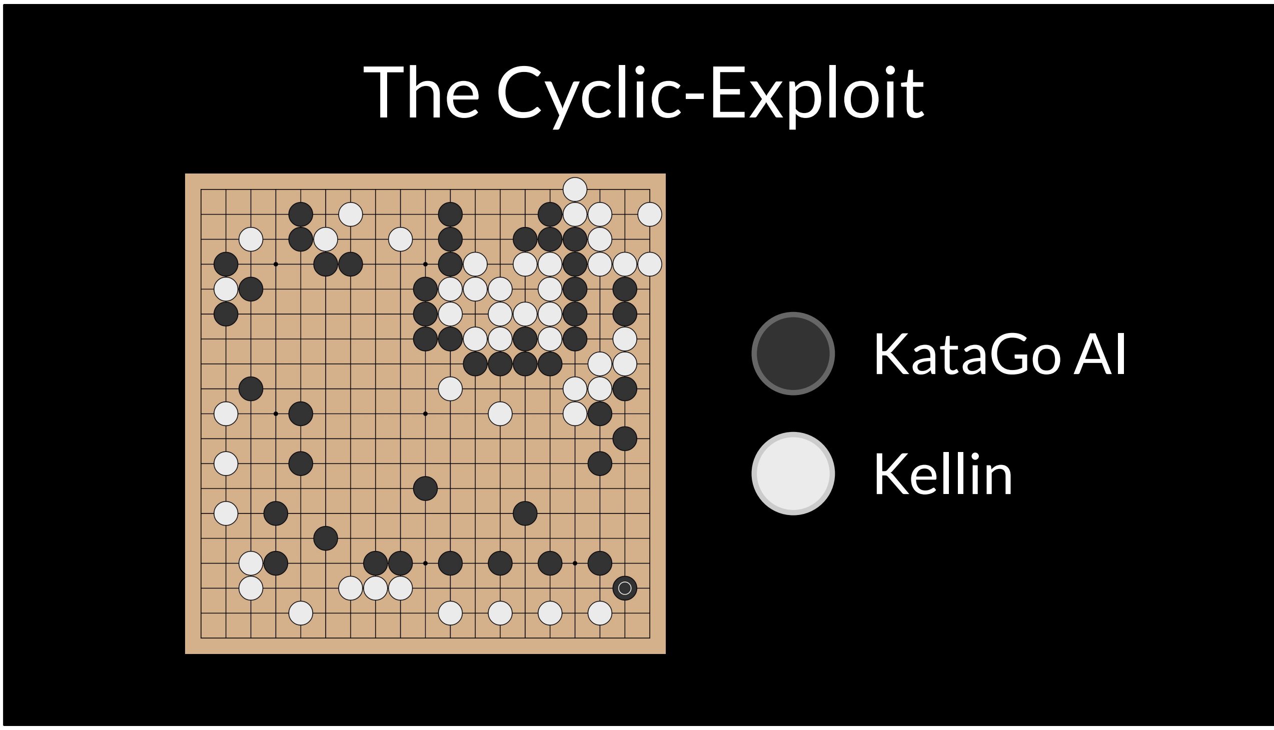


This is a losing position for Kellin at the moment. [click]

This is because black controls roughly 50 more squares than white, and in Go, whoever controls more area wins. However, Kellin has other plans. [click]

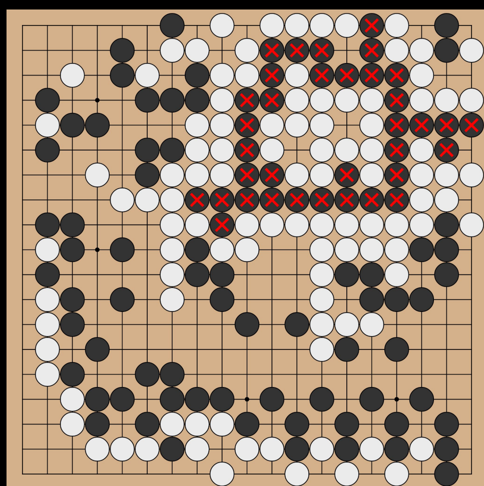


Namely, Kellin’s plan will be to kill the cyclic-group of black stones that is encircling his white group. He will do this, by slowly re-encircling the black stones from the outside. Let’s watch this in action.
[click]



[pause] KataGo, actually has numerous opportunities to stop this re-encirclement. However it does nothing. This is the hidden weakness of most modern AlphaZero-style Go AIs. When these AIs see a cyclic-group on the board, they think it is invulnerable, even when it is not. By the time KataGo realizes something is wrong, it is too late. Indeed, KataGo resigned in this position, [click, click]

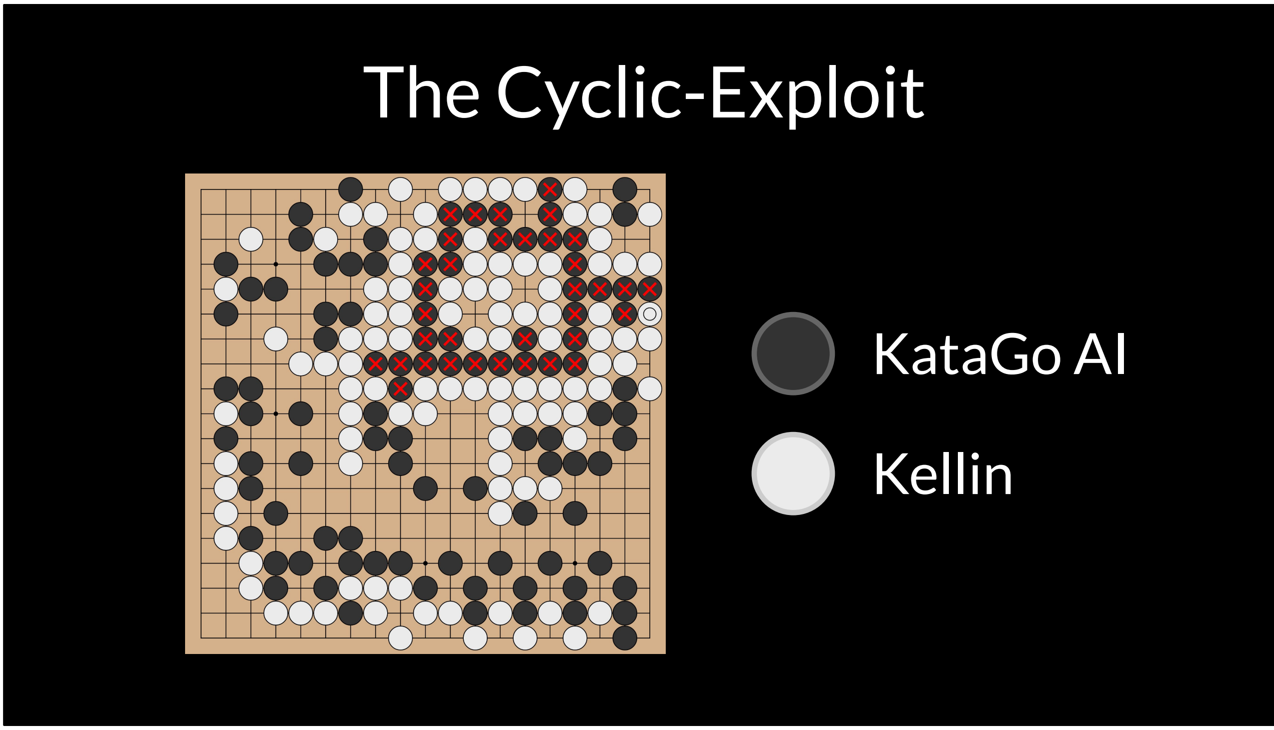
The Cyclic-Exploit



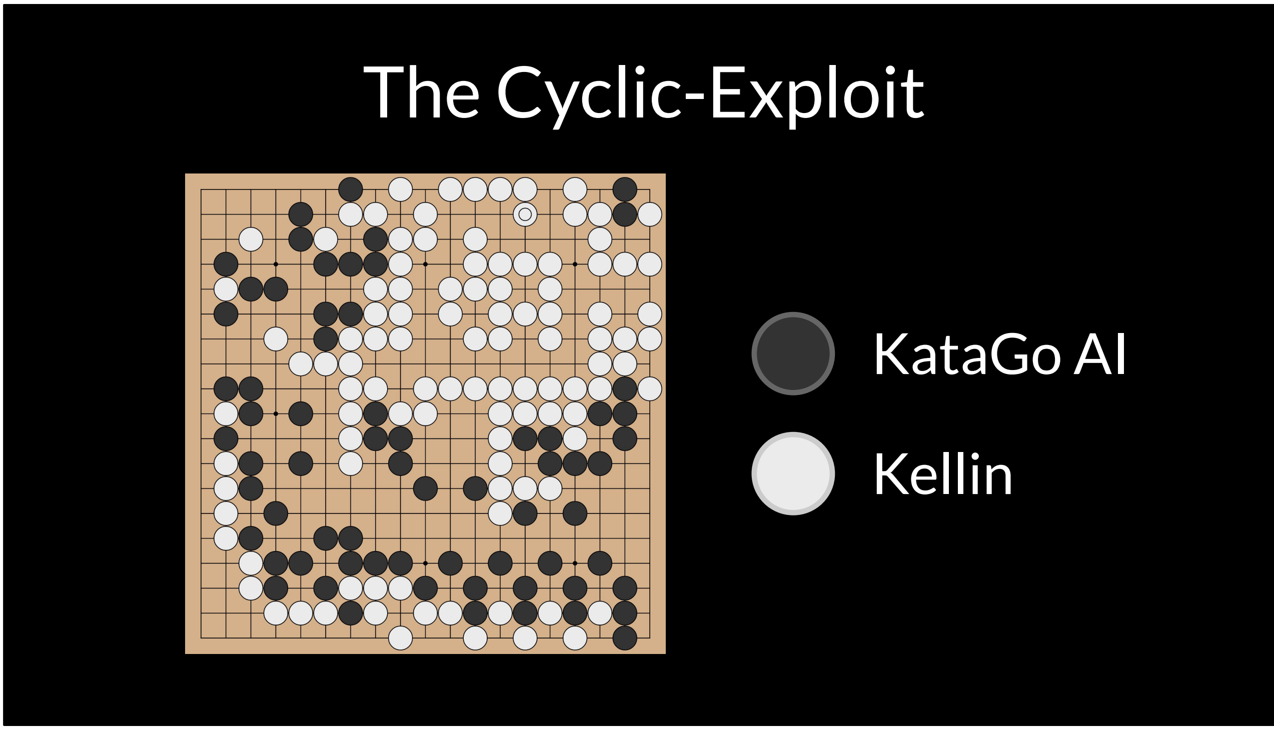
KataGo AI

Kellin

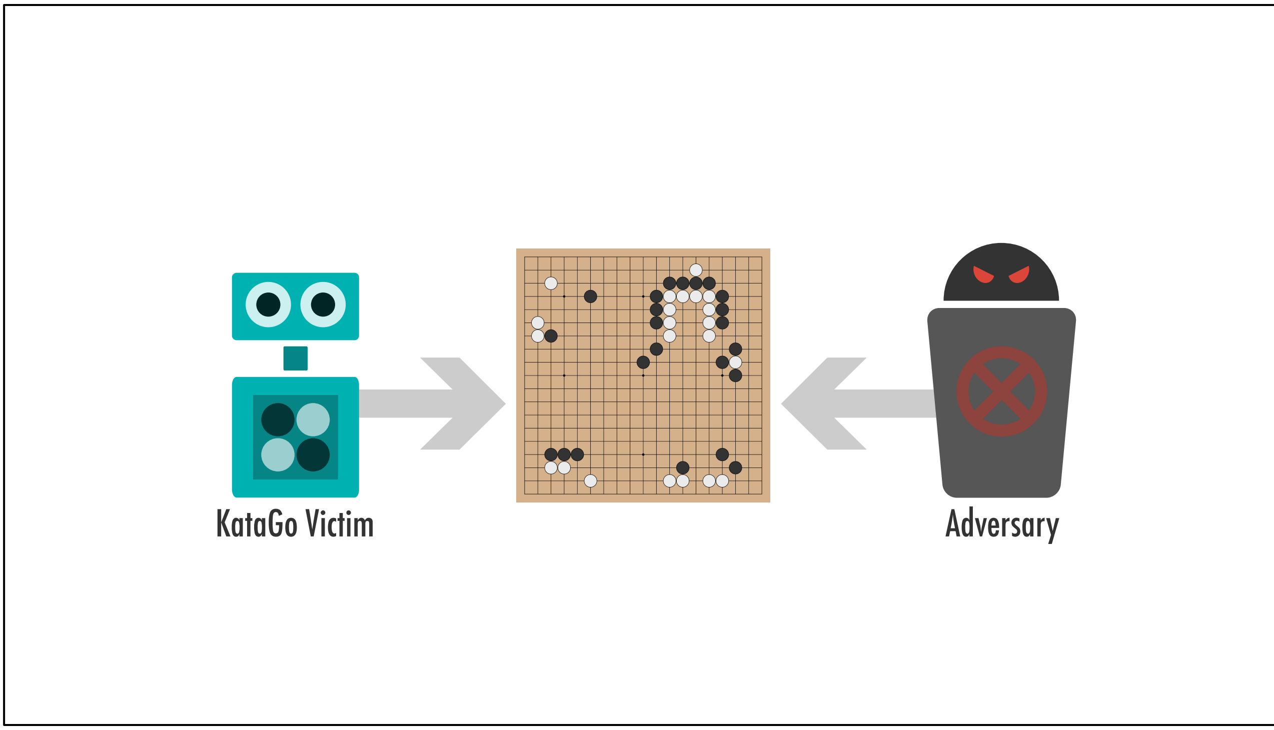
Its black cycle is guaranteed to die, as after Kellin plays on the right [click, pause]



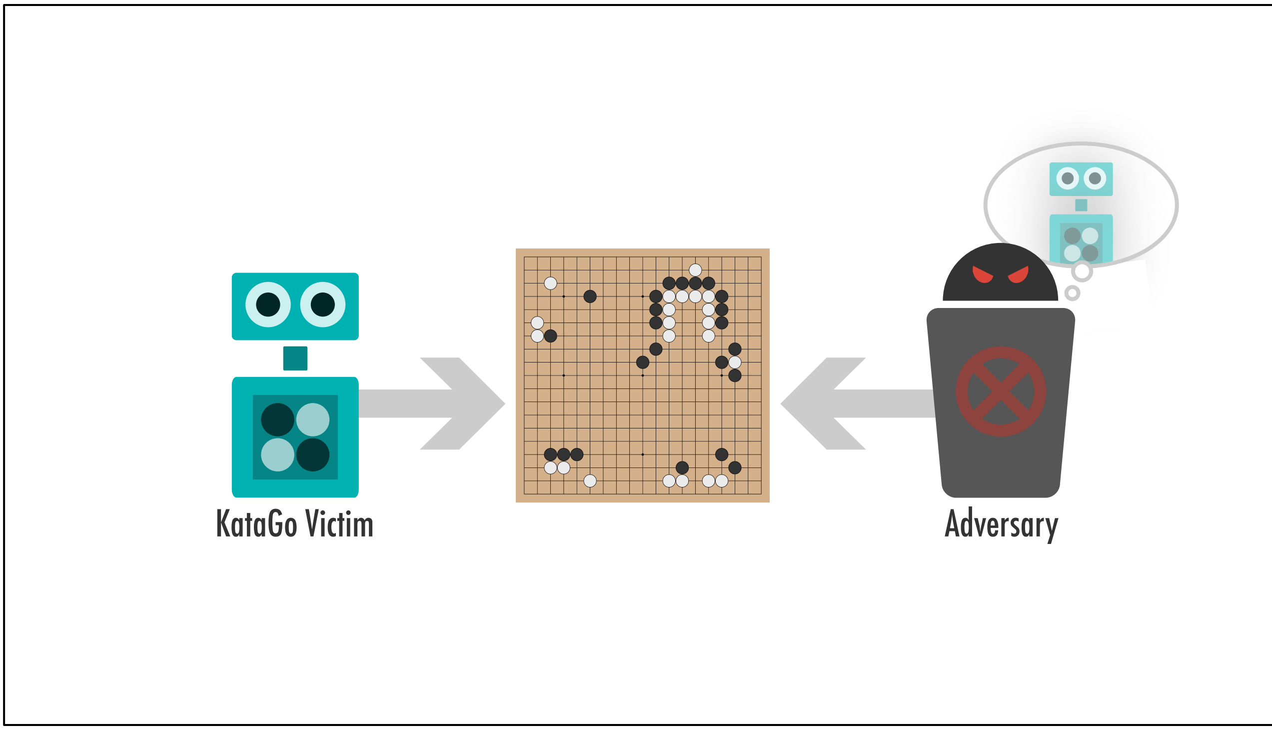
and then the top [click, pause]



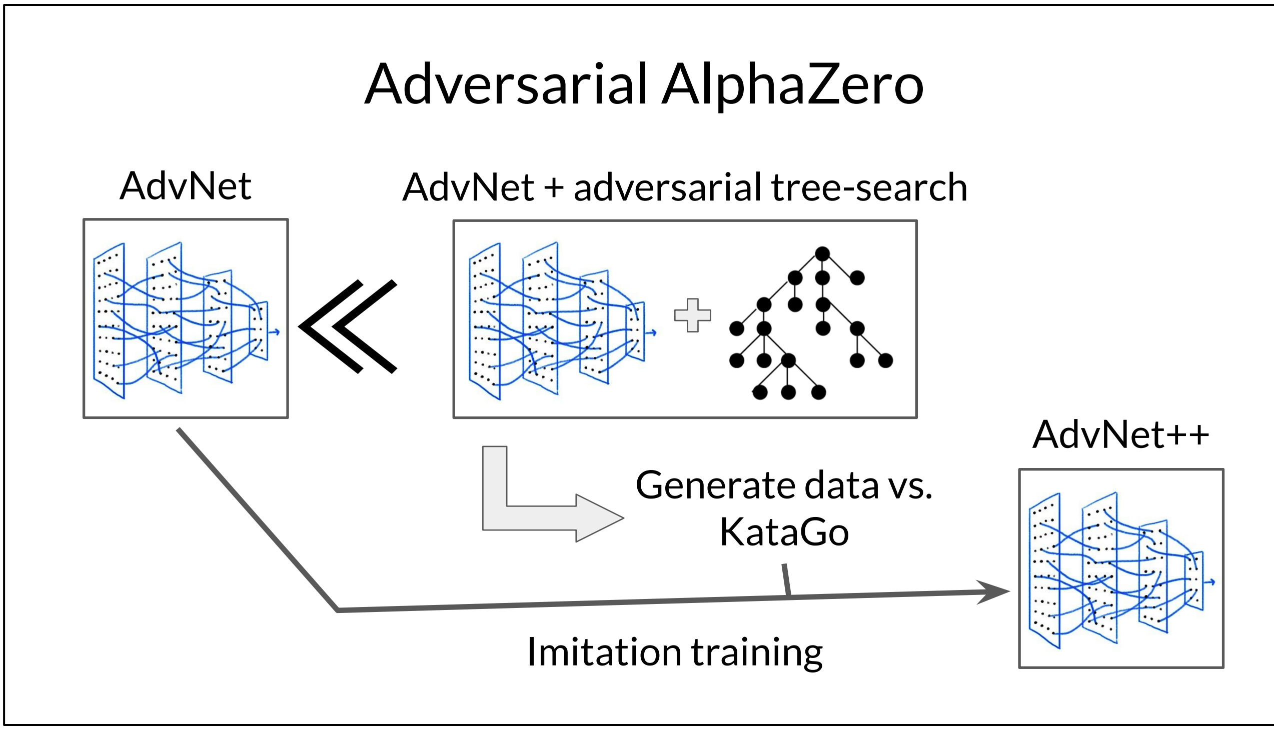
the entire group is killed and Kellin now controls the majority of the board.
[click]



How did we discover this exploit? Well, we did so by training an adversary AI to defeat KataGo. [click]



Our adversary has a special ability. Namely, it can simulate the victim's behavior when it searches over future moves.



To train our adversary, we use an adversarial variant of the AlphaZero algorithm. Here's how it works. [click] We start with a randomly initialized adversary neural network [click] Next, we augment this network with an adversarial variant of Monte-Carlo Tree Search. Tree-search is a policy improvement operator, meaning the network with tree search is a stronger adversary than the network alone. This adversarial tree-search is also where the adversary simulates possible victim responses. [click] We then pit our search-augmented adversary against KataGo, generating a dataset of behavior. [click] Finally, we train the adversary network to mimic the behavior of the search-augmented adversary. This imitation training yields a slightly stronger network.

Repeating this process, we eventually get an adversary that is able to reliably defeat KataGo via the cyclic-exploit I showed previously.

1. Superhuman performance and planning are not sufficient for robustness.

2. Adversarial optimization helps find hidden failure modes.

In summary, we showed that even superhuman Go AIs can have unexpected failure modes. Here are two key takeaways from this result. [click] The first, is that superhuman performance and planning, are not sufficient for robustness. [click] The second, is that adversarial optimization is a very useful technique for finding hidden failure modes, and should be used to improve or validate the robustness of safety-critical systems.

For more information, including many more details on our methods and results, check out our website or come find us in person at ICML 2023. Thanks for listening!