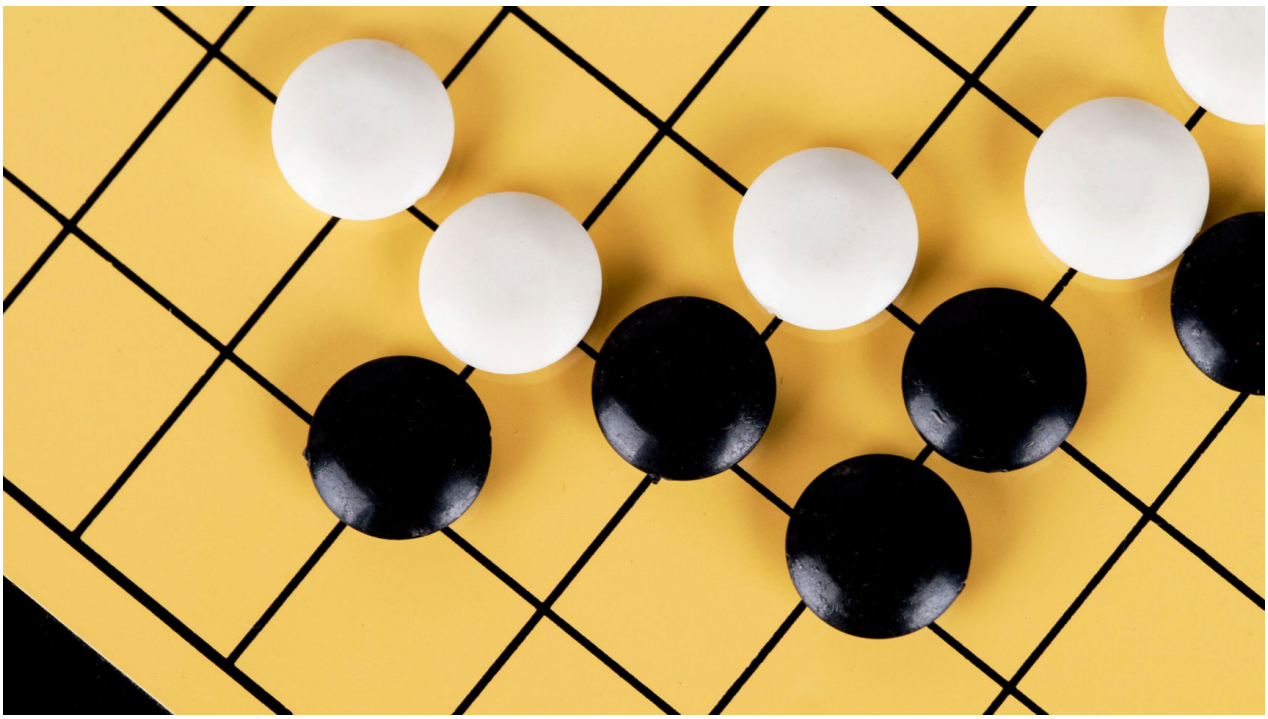
Artificial Intelligence Learns to Learn Entirely on Its Own

A new version of AlphaGo needed no human instruction to figure out how to clobber the best Go player in the world — itself.

By Kevin Hartnett



dreamdream

A mere 19 months after dethroning the world's top human Go player, the computer program AlphaGo has smashed an even more momentous barrier: It can now achieve unprecedented levels of mastery purely by teaching itself. Starting with zero knowledge of Go strategy and no training by humans, the new iteration of the program, called AlphaGo Zero, needed just three days to invent advanced strategies undiscovered by human players in the multi-millennia history of the game. By freeing artificial intelligence from a dependence on human knowledge, the breakthrough removes a primary limit on how smart machines can become.

Earlier versions of AlphaGo were taught to play the game using two methods. In the first, called supervised learning, researchers fed the program 100,000 top amateur Go games and taught it to imitate what it saw. In the second, called reinforcement learning, they had the program play itself and learn from the results.

[abstractions]AlphaGo Zero skipped the first step. The program began as a blank slate, knowing only the rules of Go, and played games against itself. At first, it placed stones randomly on the board. Over time it got better at evaluating board positions and identifying advantageous moves. It also learned many of the canonical elements of Go strategy and discovered new strategies all its own. "When you learn to imitate humans the best you can do is learn to imitate humans," said Satinder Singh, a computer scientist at the University of Michigan who was not involved with the research. "In many complex situations there are new insights you'll never discover."

After three days of training and 4.9 million training games, the researchers matched AlphaGo Zero against the earlier champion-beating version of the program. AlphaGo Zero won 100 games to zero.

To expert observers, the rout was stunning. Pure reinforcement learning would seem to be no match for the overwhelming number of possibilities in Go, which is vastly more complex than chess: You'd have expected AlphaGo Zero to spend forever searching blindly for a decent strategy. Instead, it rapidly found its way to superhuman abilities.

The efficiency of the learning process owes to a feedback loop. Like its predecessor, AlphaGo Zero determines what move to play through a process called a "tree search." The program starts with the current board and considers the possible moves. It then considers what moves its opponent could play in each of the resulting boards, and then the moves it could play in response and so on, creating a branching tree diagram that simulates different combinations of play resulting in different board setups.

AlphaGo Zero can't follow every branch of the tree all the way through, since that would require inordinate computing power. Instead, it selectively prunes branches by deciding which paths seem most promising. It makes that calculation — of which paths to prune — based on what it has learned in earlier play about the moves and overall board setups that lead to wins.

Earlier versions of AlphaGo did all this, too. What's novel about AlphaGo Zero is that instead of just running the tree search and making a move, it remembers the outcome of the tree search — and eventually of the game. It then uses that information to update its estimates of promising moves and the probability of winning from different positions. As a result, the next time it runs the tree search it can use its improved estimates, trained with the results of previous tree searches, to generate even better estimates of the best possible move.

The computational strategy that underlies AlphaGo Zero is effective primarily in situations in which you have an extremely large number of possibilities and want to find the optimal one. In the *Nature* paper describing the research, the authors of AlphaGo Zero suggest that their system could be useful in materials exploration — where you want to identify atomic combinations that yield materials with different properties — and protein folding, where you want to understand how a protein's precise three-dimensional structure determines its function.

As for Go, the effects of AlphaGo Zero are likely to be seismic. To date, gaming companies have failed in their efforts to develop world-class Go software. AlphaGo Zero is likely to change that. Andrew Jackson, executive vice president of the American Go Association, thinks it won't be long before Go apps appear on the market. This will change the way human Go players train. It will also make cheating easier.

As for AlphaGo, the future is wide open. Go is sufficiently complex that there's no telling how good a self-starting computer program can get; and AlphaGo now has a learning method to match the expansiveness of the game it was bred to play.