Jeu de go : la revanche de l’homme sur la machine

Le Point 12/04/2023

Non, l’intelligence artificielle n’est pas infaillible. L’Américain Kellin Pelrine a réussi à la piéger en recourant… à un algorithme.

*Par Hadrien Gautrot*



Kellin Pelrine montre qu'il n'y a pas de fatalité face à l'intelligence artificielle.© Kellin Pelrine

Publié le 12/04/2023 à 19h30

C'est il y a sept ans, dans un salon de l'hôtel Four Seasons de Séoul, sous le regard interloqué du monde, que l'intelligence artificielle remportait sa plus grande victoire. Lee Sedol, « meilleur joueur du monde » au go et idole d'une [Corée du Sud](https://www.lepoint.fr/tags/coree-du-sud) passionnée par ce jeu multimillénaire, [s'inclinait face à AlphaGo](https://www.lepoint.fr/high-tech-internet/l-intelligence-artificielle-ne-peut-etre-vaincue-un-maitre-de-go-abandonne-27-11-2019-2350129_47.php), un programme développé par la startup londonienne DeepMind.

S'ensuivront quatre parties, [dont une seule remportée par Lee Sedol](https://www.lepoint.fr/monde/jeu-de-go-le-champion-du-monde-bat-finalement-l-ordinateur-13-03-2016-2025031_24.php), où AlphaGo se fera remarquer non seulement pour son excellente qualité de jeu, mais également pour sa capacité à jouer des coups non conventionnels. Le champion sud-coréen parlera plus tard d'une « entité ne pouvant être défaite » au sujet de cet étrange adversaire.

Alors comment Kellin Pelrine, joueur amateur de go, a-t-il récemment réussi à battre avec plus de 90 % de réussite KataGo, cousin d'AlphaGo ayant lui aussi surclassé les meilleurs joueurs de la planète ? La clé réside dans une faiblesse encore méconnue des réseaux de neurones artificiels. En effet si ceux-ci présentent parfois des performances exceptionnelles pour une tâche précise, ils ont souvent des difficultés à généraliser cette intelligence au-delà de la tâche en question.

Figure 1. Illustration de la stratégie gagnante. En blanc, l’adversaire de KataGo – ici un algorithme – capture un groupe de pierres noires appartenant à KataGo et gagne la partie. Le groupe capturé ayant lui-même encerclé un territoire blanc plus tôt dans la partie.

Ainsi, des systèmes experts peuvent être confondus par des méthodes si simples qu'elles ne tromperaient pas un humain moyen. De fait, la stratégie employée par Kellin Pelrine consiste à laisser son adversaire encercler certaines de ses pierres, avant d'encercler progressivement ce même groupe circulaire (voir ci-contre). Une telle approche aurait été facilement contrée par un joueur amateur, mais elle a suffi à dominer une intelligence « surhumaine ».

## Un algorithme pour piéger l'IA

Pour découvrir cette stratégie, Far AI – le groupe de recherche de Kellin Pelrine – s'est aidé d'un autre algorithme conçu pour explorer les failles de KataGo. Le joueur l'a ensuite mise en pratique dans une série de duels où il n'était pas assisté par ordinateur.

La raison du succès de cette tactique d'encerclement reste matière à conjecture. Les chercheurs avancent cependant l'idée selon laquelle la rareté de cette stratégie aurait pu tromper la vigilance de l'IA. Cette dernière, ne l'ayant jamais rencontrée dans ses données d'entraînement, n'aurait pas su la repérer. Cette hypothèse semble confirmée par l'analyse de la probabilité de victoire, que KataGo calcule en temps réel à chaque instant de la partie. En effet, les chercheurs notent la grande confiance de l'IA pendant la majorité du match (situant à près de 99 % ses chances de victoire), suivie d'une chute radicale, souvent seulement un coup avant d'être capturée par encerclement.

Quelles leçons tirer de ce revirement inattendu dans le match entre l'homme et l'IA ? D'abord : que celui-ci est loin d'être plié. L'astuce et l'intuition humaines ont encore de beaux jours devant elles. Ensuite : l'essor d'une nouvelle adversité pousse l'homme à interroger ses propres limites et à conquérir de nouveaux territoires, parfois avec le concours d'outils intelligents. Enfin, si la stratégie trouvée par Kellin Pelrine devrait être facilement rendue obsolète, en l'intégrant à l'entraînement de KataGo, ce résultat pose question sur la compréhension fondamentale du jeu de go par l'IA. Sa performance ne repose-t-elle pas sur la taille phénoménale de ses données d'entraînement, plus que sur une authentique intelligence ?

Quoi qu'il en soit, ces méthodes, appelées *adversarial examples*, se déclinent dans tous les domaines touchés par l'IA : reconnaissance d'images, du son ou du langage. Leurs applications s'étendent de la protection de la vie privée au contre-espionnage, en passant par le sujet – ô combien important – de l'interprétabilité des décisions prises par ces algorithmes. La difficulté à les contourner définitivement dessine un autre futur de l'intelligence artificielle : celui, non pas d'une victoire indépassable de l'IA sur l'homme, mais d'un équilibre fragile entre des acteurs aux intérêts divergents, usant chacun d'un mélange de facultés humaines et d'outils intelligents pour maintenir leur position.