



1

ALPHAZERO APPREND TOUT SEUL

Aujourd'hui, les programmes de jeu d'échecs pour téléphones sont plus forts que Deep Blue, la machine qui, en 1997, a battu Gary Kasparov, alors champion du monde. Plus aucun espoir n'existe que des humains puissent reprendre le dessus sur les machines. C'est vrai aussi pour le jeu de go, pour lequel il a fallu toutefois attendre 2015 avant d'avoir des programmes dépassant les meilleurs joueurs.

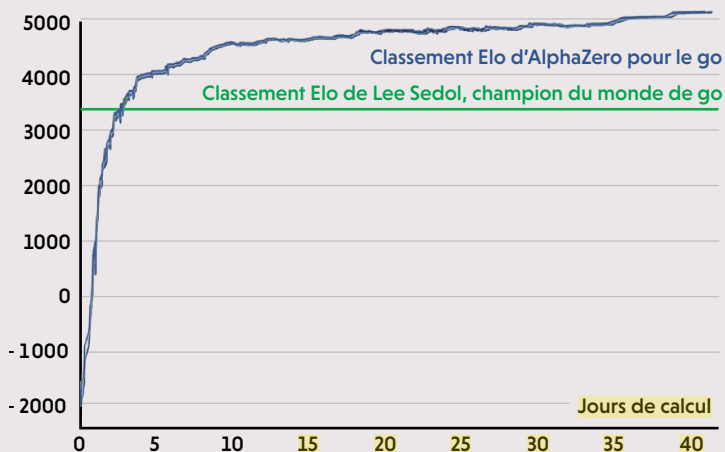
Les progrès de l'intelligence artificielle concernant ces jeux ne s'arrêtent pas là. Le programme AlphaZero, contrairement à Deep Blue et bien d'autres programmes, n'a pas besoin d'utiliser l'expertise humaine. Partant seulement des règles du jeu pour les échecs, le go et le shogi, le programme joue contre lui-même. Au départ, il joue au hasard, mais en prenant en compte les résultats des parties jouées, il apprend et se perfectionne petit à petit en utilisant un mécanisme d'apprentissage construit sur un réseau de neurones formels. Au bout d'une période de quelques heures dans le cas des échecs et du shogi, et de quelques jours pour le go, le programme atteint un niveau de jeu dépassant celui des meilleurs joueurs humains et aussi celui des meilleurs programmes. Gary Kasparov a examiné la façon dont joue le programme

d'échecs résultant de ce processus d'apprentissage par renforcement et a fait quelques commentaires intéressants :

« Contrairement aux programmes traditionnels tels que Stockfish et Fritz, qui utilisent des fonctions d'évaluation prédéfinies par des humains, ainsi que d'énormes bibliothèques d'ouvertures et de fin de partie, AlphaZero commence en ne connaissant que les règles du jeu d'échecs, sans la moindre stratégie humaine associée. En quelques heures, il joue plus de parties contre lui-même qu'on en a mémorisées dans l'histoire des échecs humains. Il apprend seul les meilleures façons de jouer, en particulier en réévaluant les pièces du jeu. Il est devenu assez puissant pour battre les meilleurs joueurs mondiaux d'échecs. Il a par exemple obtenu 28 victoires, 72 nuls et aucune défaite contre Stockfish. Je dois dire que j'ai été ravi de voir qu'AlphaZero avait un style dynamique et ouvert comme le mien. On pense en général que les machines jouent bien en choisissant des coups techniques, conduisant le plus souvent à des parties nulles. J'ai observé à l'inverse qu'AlphaZero donne la priorité à la dynamique des pièces plutôt qu'aux pièces elles-mêmes, préférant des positions qui, à mes yeux, semblaient

risquées et agressives. Les programmes reflètent généralement les priorités et les préjugés des programmeurs, mais dans le cas d'AlphaZero, qui n'est victime d'aucun préjugé, je crois que son style correspond simplement à la réalité profonde du jeu. Sa compréhension supérieure lui permet de surclasser le meilleur programme du monde et cela bien qu'il opère le calcul d'un nombre beaucoup moins important de positions par seconde. Il illustre l'adage classique : "Travaillez intelligemment, plutôt que beaucoup." »

Ainsi, AlphaZero joue Fg5 dans la position de sa partie contre Stockfish (ce qu'aucun joueur humain ni aucun programme classique ne trouvent) et gagne. Pour une discussion du coup, voir <https://youtu.be/thZzaS-noSo>. Nous avons également appris lors de son match contre Stockfish qu'AlphaZero préférerait avoir les blancs : sur les 28 victoires d'AlphaZero, 25 ont été remportées avec les pièces blanches. La machine a également compilé la fréquence à laquelle elle a utilisé les différentes ouvertures mises au point depuis des siècles par les humains, et qu'elle a retrouvées ! L'ouverture du gambit dame est finalement sa favorite.



2

ANTICIPATIONS SINGULIÈRES

Peter Diamandis est l'un des fondateurs de l'Université de la Singularité (<https://su.org>) une organisation californienne d'inspiration transhumaniste. Il a récemment interrogé divers spécialistes à propos de ce qui se passerait dans les vingt prochaines années. Il a effectué la synthèse des réponses (voir <https://www.diamandis.com/blog/countdown-to-the-singularity>)

Voici quelques-unes de ces anticipations. Précisons qu'elles ne font pas l'unanimité et qu'elles semblent plutôt optimistes.

- En 2020, des systèmes de diagnostic utilisant de l'IA et proposant des recommandations seront utilisés dans la plupart des établissements de santé américains.

- En 2020, de nouveaux catalyseurs seront découverts grâce à l'utilisation de processeurs à portes quantiques, ce qui marquera le début de la fin de la chimie traditionnelle.

- En 2022, les gens pourront légalement voyager dans des voitures autonomes partout aux États-Unis. [Si cette prédiction signifie que des voitures parfaitement autonomes existeront en 2022, elle est en contradiction avec l'avis d'autres experts, qui ne les pensent pas possibles avant une cinquantaine d'années.]

- En 2024, des missions humaines privées auront été lancées avec comme objectif la surface de Mars.

- En 2024, les premiers contrats « un centime par kilowattheure » pour l'énergie solaire et éolienne seront signés. [Aujourd'hui, les meilleurs prix dans le monde pour un kilowattheure d'électricité sont de l'ordre de 5 centimes, très rarement 2 centimes, et cela résulte d'une surproduction locale d'énergie électrique qui, sinon, serait jetée. En France, le kilowattheure est vendu autour de 15 centimes.]

- En 2030, l'IA réussira à passer le test de Turing, ce qui signifie qu'elle pourra égaler et dépasser l'intelligence humaine dans tous les domaines. [Là encore, cette prédiction semble particulièrement optimiste.]

- En 2034, les robots agiront comme des domestiques, des majordomes, des infirmières et des nourrices et deviendront des compagnons à part entière. Ils rendront possible une autonomie prolongée des personnes âgées chez elles.

- En 2036, des traitements de longévité seront disponibles et pris en charge par les assurances de santé, prolongeant ainsi la durée de vie moyenne de 30 à 40 ans.

➤ d'œuvres littéraires. La conduite automatisée de véhicules a récemment donné lieu à beaucoup de rêves, alors que les spécialistes annoncent prudemment qu'on ne fera de voitures totalement autonomes, où vous vous asseyez à l'arrière et lisez votre journal, que dans cinquante ans au mieux (voir S. Shladover, « La voiture en quête d'autonomie », *Pour la Science* n°466, août 2016).

Dans un article de cette rubrique, je mentionnais aussi les méthodes conduisant à flouer les réseaux de neurones profonds destinés à la classification d'images, ce qui conduit à ce qu'ils croient reconnaître une autruche quand on leur présente une photo où tout humain voit clairement un camion (*Pour la Science* n°488, juin 2018).

Aujourd'hui, ce domaine de recherche qui n'a pas une définition et une délimitation précises, réussit magnifiquement dans des tâches spécialisées de plus en plus nombreuses, mais rencontre deux obstacles principaux. D'une part, l'IA obtient ses résultats sans procéder comme le font les humains pour les mêmes tâches, mais en s'aidant de dispositifs techniques et de calculs dont se passent les humains : radars et lasers pour les voitures partiellement autonomes, capacités colossales de calculs et de mémorisation qui n'ont rien d'équivalent dans les cerveaux humains. D'autre part, les succès obtenus ne se combinent pas, et l'IA ne sait pas aujourd'hui former des systèmes cohérents associant les capacités spécialisées maîtrisées pour constituer des êtres autonomes doués de sens commun et dont on puisse raisonnablement soutenir qu'ils ont conscience d'eux-mêmes, possèdent une forme de sensibilité, de l'humour et des personnalités structurées et robustes.

ANNONCES PRÉMATURÉES
ET SINGULARITÉ

En 1958, Herbert Simon, qui a reçu en 1978 le prix d'économie en mémoire d'Alfred Nobel, annonçait que l'ordinateur battrait le champion humain aux échecs dans les dix années à venir. Il avait raison sur la possibilité d'un tel exploit, mais pas sur la durée d'attente : il a fallu quarante ans pour que sa prédiction se réalise.

En 1966, Irving Good, statisticien reconnu qui conseilla Stanley Kubrick pour son film 2001, *l'Odyssée de l'espace*, annonçait l'avènement vers l'an 2000 d'une superintelligence artificielle au moins équivalente à l'intelligence humaine. Elle n'est pas encore là et ceux qui prévoient sa venue ont reculé l'échéance. Raymond Kurzweil, aujourd'hui directeur de l'ingénierie chez Google, et Peter Diamandis, cofondateur et président de la Singularity University (une société privée californienne), évoquent 2030, alors que Ted et Ben Goertzel,



chercheurs en IA, pensent plutôt qu'il faudra attendre 2040, voire 2100.

Il se pourrait que l'on se trouve dans une situation analogue à celle du jeu d'échecs : on y arrivera, mais pas dans les délais annoncés, car le problème est réellement difficile, ce que l'on découvre année après année. Le plus raisonnable serait sans doute de ne pas tenter aujourd'hui de pronostics !

Liée à cette éventuelle superintelligence artificielle, l'idée d'une « singularité » à venir est évoquée sans qu'on puisse savoir vraiment de quoi il s'agit. Une date où se situerait une sorte d'asymptote verticale dans la courbe mesurant le niveau des intelligences artificielles créées sur Terre ? Un point où les machines écrasent les humains prendraient le pouvoir ? Un instant au-delà duquel il est impossible de comprendre ce qui va se passer à cause des bouleversements entraînés par l'existence d'intelligences artificielles nous dépassant ?

Tout cela n'est pas très sérieux, même en prenant en compte la célèbre « loi de Moore » (vérifiée en gros depuis cinquante ans) qui indique un doublement de capacité des dispositifs informatiques tous les dix-huit mois ou deux ans. D'une part, une fonction exponentielle comme celle envisagée par la loi de Moore n'a pas d'asymptote verticale. D'autre part, dans notre monde physique, toute croissance exponentielle rencontre rapidement des obstacles qui l'arrêtent et personne ne défend sérieusement que la loi de Moore (qui s'essouffle déjà) continuera à être valide indéfiniment. Dernier point : l'augmentation des capacités de calcul et de mémoire ne constitue pas la garantie qu'on saura les utiliser pour créer cette fameuse intelligence générale qui nous résiste.

Les annonces des Gafam (Google, Apple, Facebook, Amazon, Microsoft) sur ces sujets doivent être examinées avec méfiance. Certains spécialistes comme Jean-Gabriel Ganascia, président du comité d'éthique du CNRS et lui-même chercheur en intelligence artificielle, sont plus rigoureux et lucides que les agents publicitaires des firmes de technologie. Jean-Gabriel Ganascia considère que ceux-ci cherchent à promouvoir un projet du dépassement de la politique et des États-nations par des entreprises à l'avidité sans borne. L'idée de la singularité, utilisée pour amuser et effrayer, aurait en réalité surtout pour but d'intimider et de masquer leurs ambitions dominatrices.

Les difficultés à anticiper avec précision n'interdisent pas de s'adonner au jeu de l'anticipation en s'interrogeant sur l'apparition d'une superintelligence artificielle, quitte pour cela à imaginer plusieurs scénarios, sans donner d'échéances et en envisageant qu'éventuellement aucun ne se réalise. Jouons.

Nick Bostrom définit dans son livre une superintelligence comme tout intellect qui

excède radicalement les capacités cognitives humaines et ce dans tous les domaines. Deux hypothèses assez différentes sont envisagées.

(a) Les machines à base de silicium, y compris les ordinateurs quantiques s'ils se révèlent utiles, progressent au point que les dispositifs construits nous égalent et nous dépassent dans toutes les tâches intellectuelles possibles, cela, comme pour AlphaZero, sans que nous soyons associés directement à leur fonctionnement et sans que des éléments biologiques aient été fusionnés ou connectés à ces créatures intelligentes artificielles.

(b) Seule une association entre des éléments (humains ou non) provenant de la biologie et des dispositifs informatiques, donc un mixte de technologies inventées par l'homme et de parties utilisant ce que la vie sur Terre a patiemment élaboré, réussit ce dépassement.

ARTILECTS OU CYBORGS ?

Les êtres intelligents provenant de l'hypothèse (a) sont dénommés *artilects* (pour *artificial intellect*) par Hugo de Garis, un chercheur qui a développé un scénario pour le futur sur lequel nous allons revenir.

La science-fiction a envisagé l'hypothèse (b) depuis longtemps et dénommé cyborgs (pour *cybernetic organisms*) ces êtres hybrides superintelligents grâce au mélange et à l'imbrication de biologie et d'informatique. Imaginons par exemple que, pour suppléer à notre mémoire insuffisante, on nous greffe des puces contenant des téraoctets de données qui nous permettraient directement par la pensée de savoir tout ce qu'il y a dans l'encyclopédie Wikipédia, et que grâce à un module de calcul mathématique lui aussi accessible instantanément par la pensée, nous sachions répondre à n'importe quelle question du type « Combien fait 7^{20} ? » ou « Quelle est la dérivée cinquième de $\sin(\cos(x))$? » Notre intelligence en serait accrue !

Une troisième hypothèse est parfois évoquée, celle d'un cerveau planétaire naissant des réseaux de machines qui s'échangent des informations et constituent une sorte d'être décentralisé et pensant ; nous y reviendrons.

LE PESSIMISME DE HUGO DE GARIS

Pour Hugo de Garis, « la société se divisera en trois groupes philosophiques majeurs, agressivement opposés. Le premier groupe sera constitué par les *cosmistes* (terme dérivé du mot *cosmos*), favorables à la construction d'artilects. Le deuxième groupe sera constitué des *terrans* (dérivé du mot *Terra*, la Terre) opposés à la construction d'artilects. Le troisième groupe sera celui des *cyborgistes*, qui souhaitent devenir eux-mêmes des artilects en ajoutant des composants à leur propre cerveau humain. »

>