IA : L'erreur « artificielle » comme non-conformité à une attente « naturelle » ?

Les erreurs, déviances, écarts, biais trompeurs, sont-ils à supprimer au nom d'une fiabilité technologique et d'une neutralité axiologique, ou doivent-ils faire l'objet d'un engagement réflexif et éthique qui consisterait à choisir des traitements stratégiques et des « discriminations » éthiques ?

Que ce soit pour calculer, optimiser ou apprendre, l'IA procède par imitation et entraînement. Dans son apprentissage par renforcement, AlphaGo a appris de ses erreurs selon un processus d'auto-amélioration. Dans la partie de Go l'opposant à Lee Sedol, l'improbable 37e coup qui a surpris tout le monde était-il un coup calculé ou une erreur ? Considéré initialement comme une « erreur », il s'est avéré être un coup très judicieux, mais comme si la machine ne déterminait pas la probabilité de victoire de la même façon que l'humain. Pour l'homme, plus la marge de son territoire est importante, plus il a confiance en sa chance de gagner, tandis que pour la machine, il semble que seul le coup compte, mettant au jour une rationalité algorithmique différente, qui peut ouvrir de nouvelles perspectives stratégiques. Pourvu qu'on multiplie les données, l'IA montre une certaine fiabilité, pour un nombre de classes limité, et sans saisir toutefois le contexte. Ainsi, l'ILSVRC, qui évalue les algorithmes de traitement d'images à partir d'ImageNet, a fait baisser le taux d'erreurs de 25% à 3 % en quelques années, grâce à l'apprentissage par transfert. Cet apprentissage automatique et cette exposition aux données massives font gagner les algorithmes en fiabilité, mais, en contrepoint, ils orienteraient (« biaiseraient ») les algorithmes. C'est pourquoi on leur reproche de plus en plus des « biais » trompeurs comme des présupposés axiologiques discriminants ou des profilages tendancieux.

Or, les « biais », les erreurs et certains écarts devraient moins nous faire regretter le caractère infaillible, neutre, objectif - illusoire – que nous rappeler que les algorithmes font partie d'un processus cognitivo-technico-praxique inextricable, d'où il semble difficile de séparer l'objet et le sujet, l'artificiel et l'humain, le réel et sa représentation. L'algorithme et ses données (appréhendées cognitivement et sélectionnées) ne sont jamais neutres. On voit avec l'œil et non à travers, sans pouvoir dissocier ce qui est appréhendé de ses moyens d'appréhension. Et s'il n'y a pas de point de vue de nulle part, de « *straight view* », il n'y a pas de biais à proprement parler. Ils sont l'angle, le prisme, la focale, bref une façon de voir. Ils performent, orientent en même temps qu'ils rendent possible un mode d'apparaître ou un résultat. Autrement dit, ils conditionnent dans le sens où ils contraignent, mais également dans le sens où ils sont condition de possibilité d'un résultat. De tels prismes technico-normatifs ne sont pas des « biais » à supprimer, mais des choix à faire et à assumer. Il ne faut pas rêver : l'IA ne sera pas moins discriminante que l'homme.

Ainsi, plutôt que de vouloir chercher à rectifier une déviance « artificielle » par rapport à une norme, vouloir supprimer des « biais » trompeurs, pour tendre vers une conformité entre un résultat et un attendu, ou tendre vers un traitement partial et objectif, il conviendrait d'engager la réflexivité au cœur même de l'algorithme, c'est-à-dire au cœur des compromis et des choix. Il s'agirait de bien choisir les « biais », c'est-à-dire de choisir, à partir de différents traitements possibles, des « discriminations » éthiques, plutôt que de louvoyer sur une fiabilité technologique et une neutralité axiologique illusoires. Le choix des normes et des discriminations est une question politique.