L’IA hybride rend possible la transparence algorithmique

La transparence peut-elle s’expliquer ? Les résultats issus du Machine Learning ou du Deep Learning, sans parler des modèles eux-mêmes, sont souvent inexplicables, même pour ceux qui les programment. Dans le même temps, au-delà de la volonté politique de l’Union Européenne pour un droit à l’explication des décisions automatisées, il y a fort à parier que les citoyens que nous sommes aspireront au fur et à mesure à une plus grande transparence des algorithmes afin de comprendre, d’évaluer voire de s’opposer à toutes discriminations.
Reste à savoir ce que doit être exactement cette transparence. Doit-elle être une explication ? Une appréciation de sa complexité ? Doit-on choisir entre explicabilité et exactitude ? L’exactitude de la prédiction de certaines techniques est inversement proportionnelle à leur explicabilité.

La perspective de mieux comprendre le fonctionnement des techniques d’intelligence artificielle n’est pas désespérée. Plusieurs chercheurs travaillent à comprendre le Deep Learning en utilisant des méthodes issues de la recherche en biologie. Pour d’autres chercheurs, la recherche de l’interprétabilité est une erreur car elle empêche d’utiliser ces technologies à leur plein potentiel en les bridant aux capacités humaines.
En effet, pour être interprétable, un modèle se doit d’être simple. Mais dans ce cas, qu’est-ce que la simplicité ? Peu de facteurs explicatifs ? Une méthode basique ? Une forte discrimination ? … Le problème reste entier.

L’IA hybride ou le meilleur des deux mondes

Une des solutions consisterait à mixer intelligemment IA symbolique (système expert) au Machine Learning et Deep Learning. Les systèmes experts connurent un certain succès avant d’être dépassés par le Machine Learning. Mais l’IA symbolique conserve l’avantage de pouvoir être lisible et compréhensible par nous autres humains. Elle est fondée sur la modélisation du raisonnement logique, sur la représentation et la manipulation de la connaissance par des symboles formels. Plus prosaïquement, il s’agit de systèmes experts qui reproduisent par des règles, des décisions.

Murray Shanahan (Professor of Cognitive Robotics, Imperial College London) travaille à la création d’une IA hybride, réunissant le meilleur des deux mondes. Cela revient à « entraîner » le système, à enseigner à une autre machine les règles d’un jeu et l’état du monde qui l’entoure, afin que cette dernière puisse formuler en des termes plus abstraits ce qui est en train de se passer ».

Ce système aurait un net avantage sur le Deep Learning du fait de sa transparence et de son moindre besoin de données pour apprendre. Ces exemples montrent qu’en matière d’IA, une technologie n’est jamais complètement obsolète ou abandonnée. L’hybridation de différentes méthodes est toujours plus riche. Les évolutions législatives et sociales renforceront considérablement ce fait par l’apparition continuelle de nouvelles contraintes. Le futur de l’IA n’est

pas encore écrit.

Sans responsabilité, la transparence algorithmique est caduque

Pour autant, la transparence ne signifie pas responsabilité. La transparence algorithmique crée probablement un faux espoir. Elle cache les politiques réelles qui sont en jeu. Ouvrir le code des modèles ne suffit pas à ce que tout le monde puisse l'inspecter, ni ne permet qu'il rende des comptes. La transparence algorithmique ne mène nulle part sans considérer les données. La transparence pour la transparence n'est pas un objectif soutenable. On a besoin de la responsabilité, c'est-à-dire que les systèmes rendent des comptes de façon à pouvoir lutter, par exemple, contre toute discrimination. N'oublions pas que ces systèmes reposent sur l'expérience et que par conséquent, ils reproduiront ce qu'ils ont observés. Comment les décisions sont prises ? Sur quels critères ? Est-il plus juste de donner à chacun des chances égales ou de lutter contre l'iniquité ? Qui décide ?

Une des voies à suivre réside très certainement dans un encadrement strict de tous les traitements potentiellement discriminatoires, avec une obligation de transparence totale donnant un droit d'opposabilité. Par contre, concernant les traitements non discriminatoires, une liberté totale devrait être laissée avec pour objectif la recherche continue d'innovation et de progrès. Les utilisations qui en découlent seront les juges finaux de par la qualité des résultats, tout en respectant le consentement de chaque personne.