Moratoire artificiel

Daniel Innerarity

Cette année est en passe de devenir l'année la plus frénétique en matière d'intelligence artificielle. L'étonnement, l'enthousiasme ou la panique provoqués par ChatGPT et ses fabuleuses fonctionnalités ont été suivis par une lettre ouverte dans laquelle des scientifiques et des entrepreneurs de l'industrie technologique ont appelé à un moratoire numérique, avec autant de bonnes raisons que de procédures confuses. Ce mouvement s'est poursuivi avec l'interdiction du Chat en Italie, qui a invoqué d'éventuelles violations de la protection des données. Les appels au contrôle et à la réglementation se sont également multipliés aux États-Unis devant la Commission fédérale du commerce (FTC) pour des raisons de droit commercial, de sécurité publique et de respect de la vie privée. Un jeune homme se suicide après avoir parlé à un *chatbot* et cela semble être une nouveauté, comme si nous avions oublié, par exemple, la vague de suicides déclenchée par la lecture des *Souffrances du jeune Werther* de Goethe au xviii^{e} siècle, c'est-à-dire dans le monde analogique…

Drapeau rouge

Depuis les années 1970, l'évolution de l'intelligence artificielle suscite des vagues récurrentes de grandes espérances et de craintes apocalyptiques. En contemplant cette agitation, je me suis souvenu du *Red Flag Act*, proclamé en Angleterre en 1865 afin de prévenir les accidents dus à l'augmentation du nombre de voitures, imposant une vitesse maximale de quatre kilomètres par heure à la campagne et de six kilomètres par heure dans les villes. En outre, chaque voiture devait être précédée d'une personne à pied, munie d'un drapeau rouge pour avertir la population. Il était possible pour les hommes et les machines de se suivre à une telle vitesse, ce qui est impensable aujourd'hui compte tenu de la vitesse à laquelle nous nous

déplaçons. Il a fallu quelques années pour que l'on prenne conscience de la nature des risques et des avantages des déplacements rapides et, surtout, que la maîtrise humaine des véhicules ne dépendait pas de la limitation de leur vitesse à celle de la marche.

Il est clair que plus une technologie est sophistiquée, plus elle est performante, mais plus aussi elle comporte de risques. L'homme explore ce territoire en partie inconnu par la réflexion, qui est un moyen de mettre les processus en pause et d'anticiper les problèmes potentiels avant qu'ils ne se produisent. Dans le contexte des développements actuels de l'intelligence artificielle, des dangers tels que la discrimination, la perte de contrôle, l'insécurité de l'emploi et la désinformation semblent rendre souhaitable de ralentir autant que possible le développement technologique, afin d'avoir une approche réglementaire, de se concerter sur des critères éthiques et politiques, de mettre en place des autorités de contrôle et de certification. En supposant que la technologie n'attende pas, les auteurs de la lettre ouverte appellent à un moratoire de six mois, ce qui laisse soupçonner que ce délai est à la fois trop long et trop court, qu'il y a d'autres intérêts dans la demande, qu'il n'est pas réalisable ou qu'il comporte d'autres risques.

Un tel arrêt technologique est revendiqué au bénéfice de l'humanité dans son ensemble, mais il est clair que les gains et les pertes seraient inégalement répartis. On peut soupçonner qu'il s'agit d'une alliance des perdants, qui tireraient un certain avantage d'un ralentissement de la course technologique, mais il est également vrai qu'un tel moratoire avantagerait ceux qui disposent déjà des modèles de traitement automatique du langage naturel (*Large Language Models*, LLM) dont on veut réduire les risques. En outre, le moratoire pourrait servir précisément les institutions dont les activités sont censées être problématisées. Il est frappant de constater que parmi les premiers signataires de la charte se trouvent ceux qui n'ont jamais pris de telles mesures pour eux-mêmes, comme Elon Musk, qui a démantelé les comités d'éthique dans ses entreprises. Il est effrayant de penser que l'intelligence artificielle pourrait être dirigée à l'avenir par certains de ses signataires.

Course à la technologie

Le problème fondamental d'un moratoire est que prétendre éviter certains risques de l'intelligence artificielle en accentue d'autres : sommes-nous si

sûrs que ne pas améliorer les modèles de traitement pendant un certain temps est moins risqué que de continuer à les améliorer ? En effet, les systèmes actuels présentent de nombreux risques, mais il est également dangereux de retarder l'émergence de systèmes plus intelligents, comme le demande le moratoire. L'un de ces effets indésirables potentiels serait la perte de transparence. Si un tel moratoire était décidé, personne ne pourrait s'assurer que le travail de formation de ces modèles ne se poursuivrait pas de manière cachée. Le risque serait alors que leur développement, jusqu'alors largement ouvert et transparent, devienne de plus en plus inaccessible et opaque. Par ailleurs, que faut-il arrêter exactement, la recherche ou son application, et dans quels domaines, si ce n'est dans tous ? L'intelligence artificielle en médecine, par exemple, est une formidable occasion de sauver plus de vies ou de réduire les souffrances, ainsi que de promouvoir les économies d'énergie et de lutter contre le changement climatique qui ne ralentit pas, et ce serait donc une grave erreur d'arrêter la recherche dans ces domaines et dans d'autres. Comment faire la distinction entre ce qui peut être arrêté et ce qui ne peut pas l'être, en gardant également à l'esprit qu'il y a beaucoup de recherche fondamentale qui peut être utile dans différents domaines ?

D'autre part, une mesure aussi stricte que l'arrêt de secteurs technologiques dynamiques et compétitifs soulève de nombreux doutes quant à sa viabilité, tant pour les États que pour le secteur privé. Dans la configuration géostratégique fragmentée du monde d'aujourd'hui, où la course à la technologie est devenue l'un des principaux domaines de concurrence, une réglementation contraignante et applicable est inimaginable. Il n'y a pas non plus de raison pour que les entreprises dominantes acceptent volontairement une restriction qui pourrait mettre en péril leur position. De même, il est naïf de croire que tous les programmeurs éteindront leurs ordinateurs et que les hommes politiques du monde entier s'assiéront pendant six mois dans le but d'adopter des règles contraignantes pour tous.

Réparer le navire en haute mer

Il y a, à mon avis, une méconnaissance de la nature de la technologie, de son articulation avec l'homme et, en particulier, du potentiel de l'intelligence artificielle par rapport à l'intelligence humaine, qui est moins menacée que ne le supposent ceux qui craignent le suprémacisme

numérique. Certes, il existe un décalage de plus en plus inquiétant entre la rapidité de la technologie et la lenteur de sa régulation. Les débats politiques ou législatifs sont le plus souvent réactifs. Un moratoire aurait l'avantage de permettre l'adoption proactive du cadre réglementaire avant que la recherche ne progresse. Mais ce n'est pas ainsi que les choses fonctionnent, surtout pas avec des technologies aussi sophistiquées. L'appel au moratoire décrit un monde fictif car, d'une part, il considère que la victoire de l'intelligence artificielle sur l'intelligence humaine est possible et, d'autre part, il suggère que l'intelligence artificielle n'aurait besoin que de quelques améliorations techniques au cours d'un arrêt de développement de six mois. Comment se fait-il que la menace soit si sérieuse et que six mois de moratoire suffisent à la neutraliser ?

Si nous passons de la politique fiction à la politique réelle, nous découvrons un scénario très différent. L'Union européenne est le domaine politique où tout cela est réglementé le plus efficacement et le plus rapidement. La proposition de loi sur l'intelligence artificielle de la Commission européenne est sur la table depuis près de deux ans et les détails sont en discussion depuis lors. Même si la loi pouvait être adoptée cette année, il faudra probablement attendre encore deux ans avant qu'elle ne soit mise en œuvre dans les États membres. Plus qu'une preuve d'irresponsabilité ou de lenteur injustifiée, il s'agit d'une confirmation de la complexité de la question, à savoir qu'il n'est pas possible d'accélérer les processus réglementaires et d'arrêter le développement technologique, alors que de nombreux acteurs, y compris les secteurs technologiques à réglementer, doivent être mis d'accord.

Dans la vie réelle, les moratoires sont difficiles, partiels et d'une efficacité douteuse ; les choses se passent le plus souvent différemment. Le philosophe autrichien Otto Neurath a suggéré la métaphore de la réparation du navire en haute mer, au milieu du voyage et dans l'impossibilité de le ramener au port. Il s'agissait d'une critique de l'analogie fondationnaliste de Descartes, pour qui le modèle de la connaissance était plutôt la démolition d'un bâtiment et sa reconstruction complète. Compte tenu de la sophistication actuelle des technologies et de leur développement dans un environnement avec des acteurs très divers, il est illusoire de penser que l'évolution des technologies puisse obéir à un plan, mais cela ne doit pas nous dispenser de nous efforcer de rendre ce développement compatible avec une certaine

anticipation et la meilleure régulation possible. Il faut apprendre à utiliser les systèmes d'intelligence artificielle avec plus de prudence plutôt que ralentir la recherche. La théorie critique de la raison algorithmique doit être développée parallèlement au développement technologique et en dialogue avec ses acteurs.

C'est bien qu'elle existe

ChatGPT a surpris tout le monde, suscitant à la fois fascination et panique, en prouvant à quel point une technologie pouvait simuler des capacités humaines. Au-delà de cette première impression, il est facile de comprendre que cette technologie est moins extraordinaire qu'il n'y paraît. En effet, dans l'histoire, la plupart des techniques ont été développées pour améliorer, compléter ou même remplacer certaines activités humaines. Inventer des technologies qui font certaines choses mieux que nous n'est pas une rupture civilisationnelle, pas plus que la défaite des humains aux échecs ou au jeu de go n'a été une catastrophe. Il est important de rappeler qu'historiquement, les nouvelles technologies ont toujours provoqué des phases d'incertitude sociale, mais que celles-ci ne sont que temporaires.

La charte est un exercice d'alarmisme sur les risques hypothétiques d'une intelligence humaine de remplacement. L'idée d'un moratoire alimente les malentendus et les perceptions erronées sur l'intelligence artificielle. Elle suggère des capacités complètement exagérées des systèmes et les présente comme des outils plus puissants qu'ils ne le sont en réalité. Ce faisant, elle contribue à détourner l'attention des problèmes qui existent réellement et sur lesquels nous devons réfléchir maintenant et non dans un futur hypothétique. L'insistance constante sur des risques supposés extrêmes sert à attirer l'attention, parce que les messages extrêmes sont toujours plus intéressants que les propositions partielles et provisoires. Anticiper un certain avenir comme inexorable n'est pas scientifique et empêche de prendre des mesures concrètes importantes dans le présent pour adapter et réguler les systèmes d'intelligence artificielle.

La principale contribution de l'appel à un moratoire est de faire prendre conscience à de plus larges segments de la population qu'il y a effectivement des questions pertinentes en jeu. Ce qui est le plus précieux dans cet appel au moratoire, c'est son message performatif, à savoir attirer l'attention sur l'importance des enjeux pour la science, la technologie, l'économie, la

politique, les établissements d'enseignement et le grand public, et l'appel à forger les alliances nécessaires. Il convient donc de considérer cette déclaration comme un simple acte rhétorico-performatif visant à attirer l'attention sur un problème extrêmement urgent et important ou, pour reprendre les termes de Sundar Pichai, PDG de Google, parlant de la charte : *« c'est bien qu'elle existe »*, ni plus, ni moins.

Le problème n'est pas que l'intelligence artificielle soit trop intelligente aujourd'hui ou dans le futur, mais qu'elle sera trop peu intelligente tant que nous n'aurons pas résolu son intégration équilibrée et juste dans le monde humain et l'environnement naturel. Et cela ne se fera pas en arrêtant quoi que ce soit, mais avec plus de réflexion, de recherche, d'intelligence collective, de débat démocratique, de contrôle éthique et de régulation.