Dialogues interdisciplinaires

les risques majeurs de l'IA générative

Invité : Yoshua Bengio

Intervenantes et intervenants :

Caroline Lequesne

Hugo Loiseau

Jocelyn Maclure

Juliette Powell

Sonja Solomun

Avril 2024

obvia IVADO Mila

À propos de la série *Dialogues interdisciplinaires*

À travers une série captivante de *Dialogues interdisciplinaires* sur les impacts sociétaux de l'IA, nous convions une ou un invité et des intervenantes et intervenants, provenant des sciences et génies, de la santé et des sciences humaines et sociales, à venir discuter des avancées, des défis et des opportunités soulevés par l'IA.

Le premier dialogue de cette série débute avec Yoshua Bengio, qui, préoccupé par les développements de l'IA générative et des risques majeurs qu'ils engendrent pour la société, a initié l'organisation d'une conférence à ce sujet. Cette activité s'est déroulée le 14 août 2023 à Montréal et avait pour but d'engager une réflexion collective et interdisciplinaire sur les enjeux et risques posés par les récents développements de l'IA. La conférence a pris la forme d'un panel, animé par Juliette Powell, auquel était convié sept spécialistes provenant de disciplines variées : informatique (Yoshua Bengio et Golnoosh Farnadi), droit (Caroline Lequesne et Claire Boine), philosophie (Jocelyn Maclure), communication (Sonja Solomun) et science politique (Hugo Loiseau). Le présent document est ainsi issu de ce premier dialogue interdisciplinaire sur les impacts sociétaux de l'IA. Par la suite, les intervenantes et intervenants ont été invités à répondre de manière concise, dans la langue de leur choix, à des questions soulevées lors de cette activité[1].

Plongez-vous dans la lecture de ces conversations fascinantes, présentées sous forme de questions & réponses qui transcendent les frontières des disciplines. L'objectif est d'offrir une perspective critique et diversifiée sur l'impact de l'IA sur notre monde en constante transformation.

La série dialogues interdisciplinaires est organisée conjointement par Obvia, IVADO et Mila

L'Obvia – Observatoire international sur les impacts sociétaux de l'IA et du numérique est un réseau de recherche ouvert qui fédère les expertises de plus de 260 chercheuses et chercheurs. Au moyen d'une interrogation critique, l'Obvia a pour mission d'identifier les enjeux sociétaux de l'IA et du numérique et de contribuer à des solutions qui placent les êtres vivants et la biosphère au centre de leur cycle de développement et d'utilisation. La communauté de recherche de l'Obvia, en collaboration avec la société civile, les acteurs publics, l'industrie et les développeurs, produit des connaissances ouvertes et soutient le renforcement des capabilités individuelles et collectives.

IVADO est un consortium interdisciplinaire et intersectoriel de recherche, de formation, et de mobilisation des connaissances qui a pour mission de bâtir et de promouvoir une intelligence artificielle robuste, raisonnée et responsable. Piloté par l'Université de Montréal, avec 4 partenaires universitaires (Polytechnique Montréal, HEC Montréal, Université Laval et Université McGill), IVADO rassemble des centres de recherches, des partenaires gouvernementaux et industriels, pour coconstruire des initiatives intersectorielles ambitieuses favorisant un changement de paradigme de l'IA et de son adoption.

La communauté de Mila comprend aujourd'hui la plus grande concentration de chercheur·e·s universitaires en apprentissage profond (deep learning) au monde. L'institut se distingue par son expertise et ses innovations en langage de modélisation, en traduction automatique, en reconnaissance d'objets et en modèles génératifs. Depuis sa création, Mila oriente sa mission vers des pôles de recherche fondamentaux comme la santé, l'environnement et les changements climatiques, ainsi que l'éthique de l'IA. Mila étend son expertise et son leadership en matière d'IA pour déployer des avancées qui profiteront à toute la société. À Mila, les recherches s'effectuent en mode « science ouverte ». L'objectif : promouvoir la collaboration et favoriser le transfert de connaissances.

1 À noter que Golnoosh Farnadi et Claire Boine n'ont pu être disponibles pour cet exercice. Nous avons toutefois profité de leurs expertises et leurs points de vue lors du panel du 14 aout tenu à MILA.

Direction scientifique

Lyse Langlois
Directrice générale de l'Obvia

David Hartell
Conseiller stratégique chez IVADO

Virginie Portes
Directrice du soutien à la recherche d'IVADO

Soutien et coordination

Félix-Arnaud Morin-Bertrand
Professionnel de recherche à l'Obvia

Contributrices et contributeurs

Yoshua Bengio
Directeur scientifique de Mila et d'IVADO, professeur titulaire au Département d'informatique et de recherche opérationnelle à l'Université de Montréal

Caroline Lequesne
Maîtresse de conférences en droit public (HDR) à l'Université Côte d'Azur

Hugo Loiseau
Professeur titulaire à l'École de politique appliquée de l'Université de Sherbrooke

Jocelyn Maclure
Professeur titulaire au département de philosophie de l'Université McGill et titulaire de la chaire Stephen A. Jarislowsky sur la nature humaine et la technologie[2]

Juliette Powell
Entrepreneure, consultante et auteure dans le domaine des technologies et de l'IA

Sonja Solomun
Directrice adjointe du Centre pour les médias, la technologie et la démocratie et doctorante en communication à l'Université McGill

Produit avec le soutien financier des Fonds de recherche du Québec

Québec
Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture

ISBN : 978-2-925138-34-1
DOI : 10.61737/YRCP3187

2 Jocelyn Maclure a été aussi président de la commission de l'éthique en science et en technologie du Québec (CEST) de 2017 à février 2024.

Introduction par Lyse Langlois

« Conscience sans science et science sans conscience sont mutilées et mutilantes. »

Edgar Morin (1982/2017)

Cette première édition de *Dialogues interdisciplinaires* débute avec les questionnements sur les impacts sociétaux de Yoshua Bengio, chercheur primé à l'international en apprentissage profond[3]. Outre la reconnaissance internationale qu'il a obtenue sur le plan de sa discipline scientifique, c'est aussi un chercheur engagé socialement qui se préoccupe autant des bénéfices que des risques que peuvent avoir les transformations technologiques sur les sociétés[4]. Pour le professeur Bengio, les capacités croissantes de l'IA, notamment grâce à l'apprentissage profond, suggèrent que des niveaux d'intelligence humaine pourraient être atteints dans les deux prochaines décennies. Ce potentiel de capacité soulève des risques majeurs et des menaces pour la démocratie. Il est vrai que ces risques et dérives bénéficient d'un accélérateur efficace qu'offrent les réseaux sociaux. L'épandage de fausses nouvelles et les techniques d' hypertrucages par exemple inquiètent grandement et ont pour effet collatéraux de miner la confiance des citoyennes et citoyens. De plus en plus on observe une forme de polarisation et de fragmentation des sociétés attribué à ce phénomène.

Est-ce que la technologie dépassera les capacités humaines? Serons-nous remplacés d'ici vingt ans? Ce type de questionnements qui était aussi un autre versant des discussions soulève de grandes craintes. Un fait indéniable est à l'effet que les technologies nous ont déjà surpassées sur le plan de certaines capacités : pensons à la mémoire, à la capacité d'analyse et à traiter un grand volume de données, etc., qui en plus de réaliser avec efficacité certaines tâches, nous offre une relation personnalisée, utile et ininterrompue en répondant à des besoins qui parfois ne sont pas même pas encore formulés. C'est ce que certains appellent, la mise en place d'une puissance d'expertise grâce à l'IA dans la conquête des comportements humains. Mais au fond, que cherchons-nous à accomplir en voulant dépasser certaines capacités humaines ? Réduire l'autonomie, contrôler les activités humaines ? Quelle est la finalité poursuivie par ces différents marchés qui se pointent à l'horizon? Car sur le plan de la sensibilité, du libre-arbitre, de l'empathie, de l'imagination, un fossé demeure et n'est pas encore conquis. Est-ce que cela représente un risque d'extinction des caractéristiques humaines? Nos intervenantes et intervenants n'ont pas répondu directement à cette question qui somme toute demeure relativement spéculative. Force est d'admettre que ces discussions ont fait ressortir l'importance d'utiliser les bons termes, de mieux circonscrire la notion de risque et tout ce qui forge nos représentations pour qualifier la technologie tout en explorant plus à fond les modes de rationalités qui nous sont proposés.

On ne peut nier le fait que la technologie représente un phénomène majeur dont la portée est infiniment sociale, économique, mais aussi civilisationnelle. Malgré tout, nous avons la possibilité de reprendre le pouvoir sur la place que nous voulons bien lui accorder dans notre imaginaire collectif de même que sur les balises que nous voulons lui imposer. Il en revient somme toute à nous en tant que société de s'impliquer dans les débats et à ne pas évincer les questionnements téléologiques et normatifs face à l'inéluctabilité des avancées technologiques. Cette série a l'ambition d'ouvrir le dialogue tout en exposant différents points de vue sur des questionnements de l'heure.

« Les sciences humaines n'ont pas conscience des caractères physiques et biologiques des phénomènes humains. Les sciences naturelles n'ont pas conscience de leur inscription dans une culture, une société, une histoire. Les sciences n'ont pas conscience de leur rôle dans la société. Les sciences n'ont pas conscience des principes occultes qui commandent leurs élucidations. Les sciences n'ont pas conscience qu'il leur manque une conscience. Mais de partout naît le besoin d'une science avec conscience. Il est temps de prendre conscience de la complexité de toute réalité – physique, biologique, humaine, sociale, politique – et de la réalité de la complexité. Il est temps de prendre conscience qu'une science privée de réflexion et qu'une philosophie purement spéculative sont insuffisantes. Conscience sans science et science sans conscience sont mutilées et mutilantes. »

Edgar Morin (1982/2017)

3 Il a notamment remporté, avec Geoffrey Hinton et Yann LeCun le prix A.M Turing en 2018.

4 Depuis le début il a appuyé l'importance de se doter d'un encadrement éthique par son soutien à la Déclaration de Montréal (2018) de même qu'il a vu l'importance d'appuyer à titre de cochercheur, la création de l'Observatoire international sur les impacts sociétaux de l'IA et du numérique (Obvia) dans l'écosystème québécois. Il est aussi engagé auprès des Nations Unies en conseillant le Secrétaire général sur l'impact des technologies.

Réflexion par Yoshua Bengio[5]

Les capacités des systèmes d'IA n'ont cessé d'augmenter au cours des deux dernières décennies, souvent de manière surprenante, grâce au développement de l'apprentissage profond, pour lequel j'ai reçu le prix Turing 2018 avec mes collègues Hinton et Le Cun. Ces avancées ont conduit de nombreux chercheurs de premier plan en IA, dont nous trois, à réviser nos estimations quant au moment où les niveaux humains de compétences cognitives générales seront atteints. Alors que l'on pensait qu'il faudrait attendre des décennies, voire des siècles, moi-même et d'autres éminents scientifiques de l'IA pensent désormais que l'IA de niveau humain pourrait être développée au cours des deux prochaines décennies, voire dans les quelques années à venir. La nature des ordinateurs numériques par rapport au matériel biologique suggère que de tels niveaux de capacité pourraient alors donner aux systèmes d'IA des avantages intellectuels significatifs par rapport aux humains.

Les progrès de l'IA ont ouvert des perspectives passionnantes pour de nombreuses applications bénéfiques qui ont motivé les chercheurs comme moi tout au long de leur carrière. Ces avancées ont à juste titre attiré des investissements industriels importants et permis des progrès rapides, par exemple dans les domaines de la vision par ordinateur, du traitement du langage naturel et de la modélisation moléculaire. Cependant, elles introduisent également de nouveaux impacts négatifs et des risques pour lesquels peu d'investissements ont été faits. Ces risques sont difficiles à évaluer, mais certains pourraient être catastrophiques à l'échelle mondiale. Ils vont des menaces majeures pour la démocratie et la sécurité nationale à la possibilité de créer de nouvelles entités plus performantes que les humains, avec la perte potentielle de contrôle sur le cours de l'avenir de l'humanité.

Dans les sections suivantes, j'expliquerai comment de tels résultats catastrophiques pourraient se produire, en mettant l'accent sur quatre facteurs que les gouvernements peuvent influencer pour réduire la probabilité de tels événements. Ces facteurs sont les suivants (1) l'accès - qui peut travailler sur de puissantes IA, quels protocoles doivent être suivis, sous quel type de contrôle; (2) le désalignement - le défi de s'assurer que les IA agiront comme prévu, d'atténuer les répercussions si elles ne le font pas, et d'interdire les systèmes d'IA puissants qui ne sont pas suffisamment sûrs; (3) la puissance intellectuelle brute - les capacités d'un système d'IA, qui dépendent de la sophistication de ses algorithmes sous-jacents, des ressources informatiques et des ensembles de données sur lesquels il a été formé; et (4) la portée des actions - la capacité d'affecter le monde et de causer des dommages en dépit des moyens de défense de la société.

Il est important de noter qu'aucun des systèmes d'IA avancés actuels n'est manifestement à l'abri du risque de perte de contrôle d'une IA mal alignée. Pour minimiser ce risque ainsi que d'autres, je propose des mesures que les gouvernements peuvent prendre en s'attaquant aux quatre facteurs évoqués plus haut.

1 Premièrement, une mise en œuvre accélérée de cadres réglementaires nationaux et multilatéraux agiles et de législations qui donnent la priorité à la sécurité du public face à tous les risques et préjudices actuels et anticipés associés à l'IA, les risques plus graves nécessitant un examen plus approfondi.

2 Deuxièmement, une augmentation significative des efforts de recherche mondiaux axés sur la sécurité et la gouvernance de l'IA afin de mieux comprendre les risques existants et futurs, ainsi que d'étudier les mesures d'atténuation possibles, tant sur le plan technique que normatif. Cette recherche en science ouverte devrait se concentrer sur la sauvegarde des droits de la personne et de la démocratie, permettant la création éclairée de réglementations essentielles, de protocoles de sécurité, de méthodologies d'IA sûres et de structures de gouvernance.

5 À noter que cette réflexion de Yoshua Bengio a été rédigée en anglais et traduite en français, bien que ses autres réflexions dans la section suivante (réponses aux sept questions) aient toutes été rédigées en français. La version en langue originale de cette réflexion est disponible dans la version anglaise du document.

3 Troisièmement, investir dès maintenant dans la recherche et le développement de contre-mesures partagées et classifiées afin de protéger les citoyennes et les citoyens ainsi que la société d'éventuelles IA malveillantes ou d'acteurs mal intentionnés équipés d'IA et poursuivant des objectifs néfastes. Ces travaux devraient être menés au sein de plusieurs laboratoires hautement sécurisés et décentralisés opérant sous une supervision multilatérale, visant à minimiser les risques associés à une course aux armements en matière d'IA entre les gouvernements ou les entreprises.

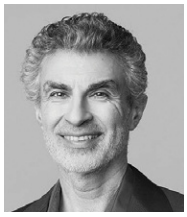
L'ampleur de ces risques est si considérable que nous devrions mobiliser nos meilleurs cerveaux et garantir des investissements majeurs dans ces efforts, au même titre que les efforts passés tels que le programme spatial ou les technologies nucléaires - afin de profiter pleinement des avantages économiques et sociaux de l'IA, tout en protégeant les sociétés, l'humanité et notre avenir commun.

Face à l'évolution rapide des technologies et à l'omniprésence grandissante de l'IA au sein de la société, il est urgent de prendre des mesures politiques. Nous ne pouvons pas nous permettre d'attendre qu'une crise - ou un événement de type « cygne noir » (faible probabilité, fort impact) se produise pour réagir. Le rythme sans précédent du développement, du déploiement et de l'adoption de l'IA exige des mesures immédiates, proactives et délibérées. Sans l'adoption rapide de mécanismes de gouvernance, je pense qu'il y a de fortes chances que les risques posés par l'IA l'emportent largement sur les possibilités d'innovation qu'elle pourrait offrir.

Q1

D'après vous, quels sont les principaux risques occasionnés par les développements de l'IA sur la société? Comment pouvons-nous les éviter?

Yoshua Bengio



Les risques qui me semblent les plus inquiétants pour la survie de la démocratie et de l'humanité pourraient être regroupés en ces trois catégories : (1) utilisations volontairement dangereuses de l'IA, (2) concentration excessive du pouvoir menaçant la démocratie, les droits humains et la stabilité géopolitique, et (3) perte de contrôle d'IAs suffisamment compétentes dans l'atteinte des objectifs qu'elles se donnent pour mettre la société et l'humanité en danger. Les horizons temporels de ces dangers se recoupent mais je les ai placés en ordre chronologique anticipé croissant. L'incertitude entourant leurs scénarios respectifs ainsi que l'amplitude de l'impact négatif anticipé vont aussi croissant, sauf pour les atteintes déjà documentés aux droits humains, faisant partie de la catégorie 2, qui pourraient être mieux mitigés si les personnes et les organisations ayant actuellement le pouvoir de décider des utilisations de l'IA étaient plus représentatives.

La catégorie (1) inclut l'utilisation de l'IA générative pour la désinformation, la fraude et la mise au point d'armes virtuelles (cyberattaques), chimiques ou biologiques. Le pire scénario pour (2) serait si une personne, une entreprise ou un gouvernement profitait d'une avance importante en IA pour s'approprier un pouvoir excessif, en commençant par la domination économique, puis politique et finalement militaire. Le scénario catastrophe le plus simple pour (3) est si, comme certains affichent déjà le vouloir, des êtres humains donnent à une IA puissante un objectif central d'auto-préservation, à l'image de l'être humain, et donc l'objectif de tout faire pour qu'on ne puisse l'arrêter, ce qui mènerait sans doute à un conflit dont l'issue serait incertaine si ces IAs sont plus compétentes que nous dans suffisamment de domaines pour être dangereuses.

Caroline Lequesne

Le débat sur les risques de l'IA emporte des considérations sur le sens de l'action publique, ses priorités et ses temporalités. Si les scénarios les plus sombres ne peuvent être ignorés, il importe toutefois de les considérer dans leurs réalités tangibles : ils sont les risques de demain et constituent l'une des trajectoires technologiques, certes possibles - voire probables -, mais à ce jour non avérées. Aussi faut-il se prémunir contre - et anticiper - ces risques, mais toutes les forces du politique ne sauraient s'y dédier. Nous y voyons là une exigence scientifique, mais aussi démocratique. Les récents débats autour de la question ont permis de mettre en évidence l'effet délétère qu'aurait une telle démarche : le sacrifice du présent au nom d'un hypothétique futur. Les atteintes aux libertés fondamentales et leurs conséquences sur les individus et les sociétés sont nombreuses, réelles et connues. De la mobilisation des IA pour manipuler l'opinion publique jusqu'aux hypertrucages pornographiques qui bouleversent l'intimité de nos existences, les chantiers sont nombreux et requièrent une action impérieuse de la part de nos dirigeants.

À défaut, les conséquences seront irrémédiables pour les démocraties comme pour les individus. Cela conduirait à envoyer un double signal : celui de l'ignorance, notamment des plus vulnérables, économiquement et socialement, premières victimes de ces systèmes; celui encore du laisser-faire aux entreprises et aux administrations qui les déploient. Il apparaît donc indispensable de construire de solides régimes de responsabilité et de redevabilité capables d'intégrer, dans le respect des libertés, l'ensemble des risques avec des réponses graduées. L'approche retenue par le législateur européen semble, à cet égard, prometteuse en ce qu'elle s'extrait de la contingence pour proposer des modes d'emploi présents et futurs.

Sonja Solomun[6]

L'un des principaux risques sociétaux posés par les développements récents de l'IA est la concentration et le déséquilibre du pouvoir entre ceux qui prennent les décisions concernant la manière dont les systèmes d'IA sont construits, déployés et gouvernés, c'est-à-dire ceux qui bénéficient de l'IA, et ceux qu'elle affecte directement. Ce déséquilibre est d'autant plus exacerbé par la centralisation de ce pouvoir au profit d'une poignée d'entreprises qui prennent actuellement des décisions qui affectent des milliards de personnes dans le monde.

Cette différence de pouvoir renvoie aux risques plus larges que les systèmes d'IA font peser sur la démocratie, c'est-à-dire deux choses essentielles : 1) la capacité de préserver et d'agir dans l'intérêt public et 2) la capacité d'influer sur la prise de décision. Les développements de l'IA doivent être intentionnellement orientés vers des résultats socialement bénéfiques qui protègent notre environnement et l'intérêt public plutôt que le développement technologique et le profit. Réciproquement, nous devons interdire l'IA lorsque les dommages causés à l'environnement et à la vie publique sont tout simplement trop importants, c'est-à-dire lorsque la surveillance conditionne l'accès à des services essentiels tels que le logement ou l'éducation, et lorsque la protection des droits humains est mise en péril.

Enfin, et c'est peut-être le point le plus important, nous tenons parfois les données pour acquises lorsque nous discutons des risques sociétaux de l'IA, en particulier en ce qui concerne les systèmes d'IA génératifs et émergents. Étant donné que ces systèmes sont entraînés sur des ensembles de données massives, nous devons garder à l'esprit les données sous-jacentes et les hypothèses sur ces données que les systèmes d'IA émettent. Les systèmes d'IA ne se contentent pas de refléter les préjugés ou les inégalités sociales existants, ils les reproduisent activement, comme l'ont montré d'innombrables études. Les travaux d'éminentes chercheuses telles que Simone Browne (2015), Wendy H.K Chun (2021), Safiya Noble (2018), Ruha Benjamin (2019) et Timnit Gebru (2020) (pour n'en citer que quelques-uns) sont incroyablement instructifs pour comprendre non seulement comment les vérités de terrain et les défauts techniques de l'IA sont enracinés dans le passé par des systèmes d'inégalité plus anciens, mais aussi, et c'est essentiel, comment nous pouvons mobiliser les résultats discriminatoires actuels pour diagnostiquer un avenir dont nous ne voulons pas.

6 Toutes les réponses de Sonja Solomun et Juliette Powell ont été rédigées en anglais et traduites en français. Les versions originale de leurs réponses sont disponibles dans le version anglaise du document.

Les modèles de changement climatique sont un exemple génératif à cet égard - lorsque les modèles mondiaux identifient l'avenir le plus probable sur la base des données actuelles et passées, nous ne modifions pas le modèle, nous essayons de changer nos actions pour éviter la « prédiction » (Chun, 2015). Au bout du compte, la concentration du pouvoir que nous observons aujourd'hui n'est pas inévitable. Éviter les risques démocratiques et imaginer de nouveaux engagements avec l'IA qui ne soient pas enracinés dans des hiérarchies de pouvoir exige que nous réfléchissions soigneusement et collectivement à la place que ces systèmes doivent occuper dans la société, ainsi qu'aux décisions que nous prenons en les utilisant. Ce n'est qu'ensuite que nous pourrons nous pencher sur les questions « du comment », notamment sur la manière dont les risques peuvent être atténués, sur la manière dont ces systèmes devraient être construits, sur les personnes qui devraient être impliquées, sur la manière dont ils devraient être supervisés et sur les personnes qui devraient les superviser. La démocratie s'effondre non pas lorsque nous sommes incapables de détecter la « vérité » à partir d'un contenu synthétique ou manipulé, mais lorsque nous perdons la capacité de débattre et de prendre des décisions concernant notre avenir collectif.

Jocelyn Maclure

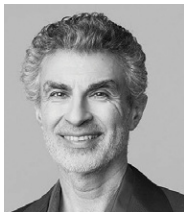
Il est utile de distinguer les risques posés par les systèmes d'IA tels qu'ils existent à l'heure actuelle de ceux créés par une éventuelle IA générale (AGI) dont les capacités cognitives seraient égales ou supérieures à celles des êtres humains. Si j'estime que les risques associés aux systèmes d'IA actuels reposant sur différents algorithmes d'apprentissage automatique sont majeurs et exigent de façon urgente le déploiement d'un cadre législatif et réglementaire robuste, les progrès récents réels en IA n'autorisent pas à penser que l'émergence d'une superintelligence artificielle radicalement autonome et cognitivement trop avancée pour être contrôlée par l'humanité est imminente, ou même vraisemblable. Soutenir cette position exigerait évidemment un argumentaire, mais il est crucial que ceux et celles qui se préoccupent des impacts de l'IA sachent que l'on peut à la fois être fortement préoccupés par l'utilisation des systèmes d'IA qui sont développés dans les laboratoires universitaires et industriels sans penser que nous avons de bonnes raisons de croire que des entités inorganiques et sans volonté propre chercheront—et seront capables—de dominer la civilisation humaine.

Parmi les risques posés par les systèmes d'IA tels qu'ils existent (et largement discutés dans le champ en ébullition de l'éthique de l'IA), on peut distinguer ceux qui sont inhérents à leur utilisation dans le but d'automatiser des processus de décision ou d'autres pratiques humaines (véhicules autonomes, production de contenus, robots sociaux, etc.) de ceux qui sont issus d'un usage malveillant de l'IA par des êtres humains. D'un côté, le recours à l'IA peut par exemple engendrer des formes de discrimination prohibées par la loi ou encore des violations du droit à la vie privée en l'absence de toute mauvaise intention. De l'autre, des humains mal intentionnés par exemple peuvent utiliser l'IA générative pour désinformer et polluer l'espace public. C'est pour atténuer ces risques et responsabiliser tous les acteurs impliqués dans le développement et le déploiement de l'IA qu'un cadre réglementaire contraignant doit être mis en œuvre.

Q2

Que faire pour que le développement de l'IA profite à tous et éviter la concentration du pouvoir?

Yoshua Bengio



La gravité de cet enjeu va augmenter au fur et à mesure que les capacités de l'IA vont avancer. On peut douter des mécanismes de marché pour empêcher la concentration du pouvoir découlant d'IA très puissantes puisque l'objectif d'une entreprise est d'établir autant que possible une domination économique, et que l'argent achète le pouvoir. La compétition entre entreprises peut contrebalancer cela, mais les coûts de calcul énormes associés aux IA de pointe et le talent très spécialisé requis mènent comme on le voit déjà à la concentration de ce pouvoir entre un très petit nombre d'entreprises.

Il est donc important que des organisations qui ne soient pas à but lucratif mais plutôt dont la mission est le bien commun soient à même (a) d'évaluer les risques associés aux IA d'autres organisations, par exemple via des audits mandatés par la réglementation et (b) d'offrir une alternative aux entreprises d'IA dont les objectifs soient d'une part la protection du public et d'autre part l'application de l'IA pour le bien général, par exemple les ODD de l'ONU, selon une gouvernance inclusive et démocratiquement établie donnant une voix à tous, incluant la défense des intérêts des pays n'ayant pas accès à un tel pouvoir.

Sonja Solomun

Il s'agit d'une question cruciale car elle présuppose une distinction importante : il y a une différence entre veiller à ce que des résultats « néfastes » ne se produisent pas (ce que de nombreux cadres réglementaires, universitaires et issus la société civile prennent pour base) et veiller à ce que des choses « bénéfiques » se produisent, en particulier pour celles et ceux qui risquent de subir les effets les plus néfastes de l'IA. Il ne suffit plus de minimiser les effets néfastes de l'IA, car ce qui reste n'est jamais un « outil » neutre, mais plutôt un système sociotechnique dynamique et évolutif qui émet implicitement ou explicitement des affirmations politiques en fonction de l'endroit et de la manière dont il est utilisé dans la société (en d'autres termes, cela ne résoudrait pas le « problème de désalignement » dont Yoshua Bengio a parlé plus tôt). Aller au-delà des approches de réduction des méfaits nous permet également de poser des questions collectives sur le bien-fondé d'un cas d'utilisation particulier ou sur l'existence de cadres alternatifs.

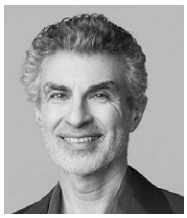
L'une des façons d'orienter le développement de l'IA dans l'intérêt public et de s'assurer que nous évaluons ces systèmes en fonction de leur impact social est de mettre en place une réglementation indépendante qui vise intentionnellement à orienter les résultats vers le bien commun. La gouvernance de l'IA a besoin d'une contribution collective et continue d'une pluralité de voix, en particulier des communautés concernées (par le biais d'éventuels comités de révision, de l'établissement de normes et d'autres mécanismes, y compris l'application de la loi).

Les cadres réglementaires de l'IA doivent également être transparents en ce qui concerne la documentation des modèles et des données et, ce qui est peut-être le plus important, des mécanismes de responsabilisation robustes, y compris un régime de responsabilité, des droits solides en matière de données avec des catégories et des interdictions clairement définies pour les risques élevés, ainsi que des cadres qui abordent le déploiement de l'IA et les risques au-delà des droits et des préjudices individuels pour y inclure les risques systémiques et environnementaux.

Q3

Comment préserver les valeurs démocratiques face à l'utilisation malveillante de l'IA?

Yoshua Bengio



Il y a plusieurs sources d'inquiétude concernant les valeurs et institutions démocratiques avec la montée en puissance de l'IA. D'une part, l'utilisation malveillante de l'IA pourrait mener à un accroissement de la surveillance étatique, aussi aidée par l'IA, dans le but annoncé de protéger la population. Cependant cela pourrait aussi justifier et permettre à des gouvernements avec des tendances autoritaires d'augmenter et centraliser leur pouvoir grâce à l'IA, et donc à menacer la démocratie. Dans le cas d'utilisation malveillante de l'IA par d'autres États et visant nos démocraties, cela pourrait mener à une course aux armements basés sur l'IA, une situation dangereuse à la fois pour la stabilité géopolitique et concernant des erreurs pouvant mener à une perte de contrôle de l'IA.

Il me semble donc essentiel de mettre en place une gouvernance démocratique et internationale de l'IA qui interdirait son développement à des fins de domination militaire ou politique. Les organisations développant des IAs de pointe devraient d'abord et avant tout servir le bien public, les institutions démocratiques, les objectifs de développement soutenable de l'ONU et la défense contre des utilisations malveillantes de l'IA.

Caroline Lequesne

La question constitue en elle-même un défi en ce qu'elle confronte deux dynamiques inconciliables par nature : détruire par l'IA ou construire ensemble. Tout l'enjeu est de ne pas sacrifier le second au premier; de ne pas laisser la peur paralyser des processus vertueux de partage, de débats, et de droits fondamentaux. Nous sommes conscients que la démocratie n'apportera pas toutes les réponses escomptées. Cet état de droit et de fait ne doit pas pour autant conduire à renoncer à la démocratisation de l'IA. Celle-ci repose sur deux éléments fondamentaux : un encadrement conforme aux libertés fondamentales d'une part; l'implication citoyenne de l'autre.

Le premier n'est pas sans rappeler les débats des années 2000 face à l'expansion de la sphère du marché et la financiarisation de la société : annonçaient-elles la fin de la démocratie? Les travaux de nombreux économistes, au premier rang desquelles Jean-Paul Fitoussi, permirent d'établir que le marché ne survivrait pas, à terme, sans démocratie. Il en est de même pour l'innovation technologique : le droit de l'IA n'a pas vocation à être un droit de l'exceptionnalité et du secret, et doit s'intégrer à un régime de liberté. De surcroît, il apparaît fondamental que ce droit soit encore le droit de tous.

Pour ce faire, le citoyen jusqu'alors écarté au nom de considérations techniques - ou d'enjeux économiques et sécuritaires impérieux - a été largement évincé des débats. Or, comment faire société sans l'engagement du plus grand nombre? Il apparaît à cet égard indispensable d'institutionnaliser un espace public technologique à côté des laboratoires. Les travaux sociologiques sur le nucléaire témoignent de la complémentarité vertueuse entre homme de science et citoyens. Cette citoyenneté technologique reste à penser et construire.

Jocelyn Maclure[7]

Cette réponse déplaira aux libertariens de droite et de gauche, mais la seule approche réaliste dans le monde dans lequel nous vivons est celle d'un interventionnisme étatique robuste : réglementation, régulation et redistribution. Les lois sectorielles existantes ont été ou sont présentement en cours de révision afin de demeurer pertinentes et efficaces à l'ère de l'IA et du numérique (lois sur la vie privée, la concurrence, la protection des consommateurs, le droit d'auteur, etc.). Des lois encadrant spécifiquement le développement et l'usage de l'IA doivent être adoptées (AI ACT, LIAD, etc.).

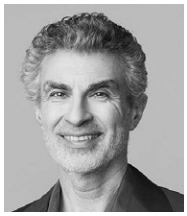
Différents modes de taxation des revenus des grandes entreprises du web doivent être déployés afin de soutenir les écosystèmes durement touchés par la révolution numérique (médias, artistes, etc.) et, de façon plus générale, de permettre aux États de s'acquitter de leurs responsabilités en matière de redistribution de la richesse et de réduire les fractures numériques. On peut bien entendu souhaiter une plus grande démocratisation des entreprises et des milieux de travail, mais cette transformation ne se fera pas de façon spontanée; elle exigera des changements législatifs.

7 Jocelyn Maclure a indiqué qu'il répondait simultanément aux questions 3 et 4.

Q4

Quelles sont les mesures à mettre en place pour protéger la société en matière de cybersécurité (utilisation malveillante de l'IA) ?

Yoshua Bengio



La mesure la plus importante serait que les très gros modèles de langue capables de programmer au-delà d'un certain niveau seuil doivent être enregistrés auprès du gouvernement, ne doivent pas être distribués sur internet, accessibles sur internet ou pas suffisamment protégés des cyberattaques pour en copier le code et les paramètres. De plus, on devrait s'assurer de retirer des données d'entraînement les codes qui sont utilisés pour des cyberattaques. Deuxièmement, on devrait mettre en place des accords internationaux pour que de telles mesures soient appliquées dans le monde entier. Par ailleurs, on devrait investir dans un réseau international de laboratoires à but non lucratif mais financés par le public et dont l'objectif serait de mettre en place des mesures défensives (incluant la surveillance de cyberattaques autrement silencieuses) en utilisant des IAs sécuritaires (ce qui en soit demande aussi des investissements en recherche).

Hugo Loiseau

La géopolitique de la cybersécurité a ceci de particulier que c'est un domaine qui augmente sa surface d'attaque proportionnellement avec le développement de son arsenal. Il en sera de même pour l'IA. Or, la question qui se pose est celle du contrôle de cet arsenal. Les études en cybersécurité nous indiquent que la non-prolifération des cyberarmes ne fonctionne guère en ce moment, car les États sont dans une logique de suprématie de la puissance dans le cyberespace qui relève d'un dilemme de la sécurité plutôt classique. De plus, les États agissent selon un principe d'*internal balancing* (Waltz, 1979) où une innovation qui donne un avantage en termes de puissance à un État sera copiée par les autres États.

Dans cette logique appliquée à l'IA, les États les plus puissants voudront contrôler ce que font les autres États tout en se laissant la liberté de développer eux-mêmes leurs propres arsenaux à base d'IA selon leurs propres intérêts. À cet égard, le texte de Joseph Nye (2011) sur ce que peut retenir la cybersécurité des enseignements du nucléaire comme technologie duale s'applique grandement à l'IA. En ce sens, les besoins de gouvernance sont immenses pour l'IA (voir la question suivante).

Juliette Powell[8]

Tout comme les ingénieures et ingénieurs mécaniques et électriques sont tenus responsables par leur profession, les ingénieures et ingénieurs informatiques et de systèmes doivent être formés et reformés / certifiés et recertifiés en matière d'IA responsable afin que si/quand des humains sont lésés par le déploiement et/ou l'utilisation de ces technologies, elles et ils soient tenus responsables de leurs créations, y compris lorsqu'ils laissent des failles susceptibles d'être piratées. Traditionnellement, les « opérateurs » sont tenus pour responsables, comme dans le cas de l'accident de la voiture autopilotée d'Uber en 2018 à Tempe, en Arizona, qui a tué une femme traversant la rue avec son vélo.

En ce qui concerne les accords internationaux, ils devront être vigoureusement appliqués et il n'est pas évident de savoir comment ils peuvent l'être tout en privilégiant l'innovation et la compétitivité.

De ce point de vue, un réseau mondial de laboratoires à but non lucratif pourrait s'avérer nécessaire pour fournir des preuves indépendantes des dommages potentiels causés aux humains. Ces nouvelles institutions mondiales exigeraient une représentation égale, incluant un grand nombre de personnes en dehors du domaine de l'IA. Ces équipes interdisciplinaires qui maîtrisent la sociologie et la psychologie, la philosophie, l'anthropologie et le droit seraient essentielles pour traiter ces très grandes questions au nom de l'humanité, en plus des autres logiques de pouvoir qui entourent l'IA : entreprises, ingénierie, gouvernement et milieu académique.

8 Voir la note de bas de page numéro 6.

Q5

À quel niveau devrions-nous instaurer une gouvernance de l'IA? (local, régional, national ou international)?

Hugo Loiseau

Dans le domaine de l'IA, la coopération internationale et le droit international sont en retard par rapport aux risques et menaces qui lui sont associés. Ainsi donc, les besoins de gouvernance sont très importants, et ce à tous les niveaux. Local et municipal, notamment pour l'acceptabilité sociale de l'IA et de ses infrastructures énergivores et consommatrices d'eau pour le refroidissement. Régional pour tous les enjeux de gouvernance transfrontalière (échanges commerciaux, chaînes de valeurs, gestion des flux...). National pour les politiques de recherches et développement, mais aussi pour l'encadrement de l'IA dans un ordre juridique interne qui oriente les niveaux locaux, régionaux et internationaux. International, enfin, pour l'élaboration d'un droit international de l'IA visant à réguler, autant que faire se peut, les centres de recherche et les entreprises (du numérique ou autres).

Sur le plan international, le modèle à suivre, à mon avis, serait celui du régime de non-prolifération nucléaire. Loin d'être parfait, ce régime international a néanmoins permis l'émergence de normes internationales mises en œuvre par l'Agence internationale de l'énergie atomique (AIEA) à travers le Traité sur la non-prolifération des armes nucléaires (TNP). Cette agence relevant de l'ONU promeut l'usage pacifique de l'atome tout en limitant ses usages militaires.

Caroline Lequesne

La question appelle à des distinctions entre être et devoir être; entre ce qui s'apparente à l'espoir bien-fondé et l'ordre de l'action pragmatique. Si les deux niveaux de réponses ne sont pas irréconciliables – et se nourrissent - le premier invite à s'affranchir du second pour être « ambitieux ». Il serait ainsi souhaitable, au regard des enjeux économiques, démocratiques et civilisationnels identifiés, qu'une réponse soit apportée à l'échelon international. Un accord emportant l'assentiment du plus grand nombre, visant à réduire la course à la concurrence et ses effets au profit d'un intérêt général, commun et humain constituerait si ce n'est une panacée, le symbole d'une prise de conscience collective indispensable. Pourtant, l'expérience des « trente globales » concernant les grands défis planétaires témoigne de la faiblesse et des limites du droit international.

On conviendra parallèlement que les initiatives nationales semblent bien modestes, voire inadaptées. Aussi, d'un point de vue pragmatique, l'échelon régional apparaît comme le plus prometteur. L'Europe trace à cet égard une voie : si les compromis demeurent complexes à construire, le niveau d'exigences y est accru et la mise en œuvre de règles, au sein d'un large marché de consommateurs, susceptible de produire un effet d'entraînement sur les autres zones (où l'on reparler du « Brussel effect », théorisé dans des domaines corollaires tels que la protection des données à caractère personnel).

Juliette Powell

D'un point de vue mondial, trois entreprises : Microsoft, Google (Alphabet) et Amazon se disputent la place de premier fournisseur mondial d'IA dans le cloud. Dans un secteur qui connaît une croissance de 70 % ou plus chaque année, la capacité des réglementations à contraindre les comportements est difficile. De plus, l'utilisation des LLMs peut sembler être un changement assez mineur dans notre façon de fonctionner, mais leur impact pourrait changer le monde.

C'est pourquoi nous devons mettre en place une gouvernance de l'IA à tous les niveaux, en commençant par le niveau individuel. En tant que parties prenantes de l'écosystème de l'IA, nous avons la responsabilité de faire preuve d'esprit critique lors de l'examen et de l'expérimentation des technologies émergentes. Nous nous devons également, à tous les niveaux, de développer un calcul du risque intentionnel autour de l'IA. Les fonctions de risque peuvent estimer les changements extrêmes, mais ils ne sont pas aléatoires. Pour les organisations et les gouvernements, les gains ou les pertes peuvent généralement être estimés - et avec une connaissance suffisamment approfondie d'un processus ou d'un événement particulier, une estimation peut être raisonnablement fiable. Par exemple, les banques utilisent des fonctions de risque pour évaluer la solvabilité; une personne ayant une bonne cote de crédit représente un meilleur risque qu'une personne ayant une mauvaise cote de crédit et peu de ressources. Pour établir un calcul du risque intentionnel, un décideur pourrait commencer par reconnaître les conséquences potentielles importantes de chaque projet de système d'IA. Pour chaque conséquence, il pourrait créer une matrice, en comparant la gravité de l'impact à la probabilité d'occurrence; établir une valeur de risque pour chaque cellule de la matrice, sur la base de ces matrices; créer des scénarios pour la combinaison des risques, qui donnent une idée précise de l'éventail des résultats; prendre une décision sur la base de ces scénarios, en établissant des moyens de reconnaître rapidement les imprévus ou les anomalies, afin que l'entreprise puisse rapidement changer d'orientation.

Au niveau local, régional et national, les entreprises ont souvent plus de pouvoir que les gouvernements pour effectuer des changements rapides. Les entreprises et les gouvernements doivent donc travailler ensemble, ainsi qu'avec les autres parties prenantes de l'écosystème de l'IA, et ce de manière égalitaire et sans rapport de force, afin que tous puissent s'exprimer et être représentés dans les initiatives de gouvernance

Sonja Solomun

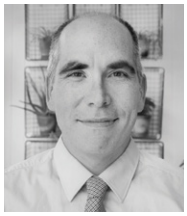
Si je me rallie à l'appel à la gouvernance à tous les niveaux, une réglementation nationale solide est cruciale et devrait idéalement être mise en œuvre par un régulateur indépendant doté de solides mécanismes de supervision et d'application. Une option est celle du modèle basé sur les risques de l'Union européenne, qui consiste à réglementer les cas d'utilisation plutôt qu'une technologie donnée « en soi ». Bien que l'approche fondée sur les risques ne soit pas sans défaut, il s'agit d'un point de départ pour encadrer le débat réglementaire et passer de la simple atténuation des effets néfastes à des mesures plus proactives - avant qu'un préjudice potentiel ou démontrable ne se produise. Les évaluations des risques et les audits indépendants (en particulier avant la collecte des données et le déploiement des systèmes d'IA) seraient des mécanismes de responsabilisation essentiels à suivre.

Cependant, la minimisation des risques liés aux systèmes d'IA n'est qu'une solution réglementaire partielle, étant donné la persistance des préjudices malgré un processus rigoureux d'évaluation, de test et d'atténuation des risques. C'est particulièrement le cas pour les préjudices systémiques tels que le coût énergétique des grands modèles de langage, par exemple. Au-delà de la réduction des risques, la réglementation de l'IA devrait être élargie au-delà des conceptualisations individuelles du risque, des droits et de la sécurité pour inclure les risques systémiques et environnementaux (Ada Lovelace Institute, 2023). Pour commencer, la réglementation devrait avoir pour objectif d'orienter le développement et l'utilisation de l'IA vers l'intérêt public, ce qui pourrait être la tâche d'un organe de gouvernance mondial.

Dans le même ordre d'idées, il est urgent de renforcer la coordination internationale de la gouvernance de l'IA, d'autant plus que les pays ont des lois et des priorités nationales distinctes. Les démocraties du monde entier doivent tirer leurs propres leviers politiques et réglementaires en direction d'un cadre commun de responsabilité et de contrôle robustes des systèmes d'IA. Pour ce faire, il faudra concrétiser les modèles de gouvernance internationale de l'IA au-delà des valeurs et des principes, afin de fixer des engagements formels ancrés dans les droits de l'homme. Il est essentiel qu'un organe mondial de gouvernance de l'IA dispose de pouvoirs d'exécution allant au-delà de simples conseils en matière d'engagements.

La gouvernance mondiale de l'IA devrait également, dans l'idéal, s'orienter vers des paramètres communs, y compris la justice environnementale, et inclure des interdictions spécifiques strictes pour les systèmes d'IA dangereux ou l'utilisation dans des domaines à haut risque. Il est important que les pays « du Sud » et les pays « en développement » soient inclus dans la gouvernance mondiale de l'IA afin que le paysage réglementaire ne soit pas dominé par les valeurs et les marchés européens et nord-américains.

Jocelyn Maclure



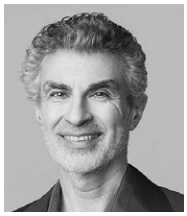
Si une gouvernance internationale des systèmes d'IA est urgente et nécessaire, il faut résister à la tentation de penser qu'il est vain d'encadrer l'IA aux autres échelles de la gouvernance politique (urbaine, régionale, nationale, continentale). Les États démocratiques ont la responsabilité d'assurer le respect des droits fondamentaux et énoncent les règles déterminant le fonctionnement des institutions et des sphères d'activité névralgiques eu égard au déploiement de l'IA (système de santé, processus électoraux, pouvoir judiciaire, transports, télécommunications, éducation, etc.).

L'édification d'un droit international contraignant est hautement souhaitable afin de favoriser un nivellement par le haut des normes et des standards encadrant le développement et la commercialisation des systèmes d'IA, mais il serait dramatique que les États et gouvernements locaux n'élaborent pas à courte échéance des modes de gouvernance robustes de l'IA. Par ailleurs, bien qu'il s'agisse d'un texte non contraignant, il est à noter que la Recommandation sur l'éthique de l'intelligence artificielle de l'UNESCO a été adoptée par les 193 États membres de l'organisation à la suite de négociations multilatérales étendues.

Q6

Est-ce que les développements de l'IA redéfinissent les propriétés de l'intelligence humaine? Pouvons-nous être "surpassé.e.s" par une IA?

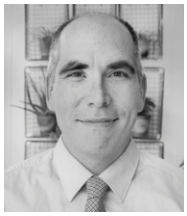
Yoshua Bengio



La recherche en IA nous apprend que l'intelligence est multiple (une entité peut être intellectuellement compétente pour certaines choses et pas d'autres) et peut être assez différente de la nôtre (ce que l'éthologie nous disait aussi) même quand l'intelligence humaine a été une source majeure d'inspiration. C'est à la fois excitant, parce que nous pourrions apprendre beaucoup d'autres formes d'intelligence, mais aussi inquiétant parce que des IA futures pourraient interpréter le monde, et en particulier la moralité des actions, d'une manière très différente de nous mais aussi potentiellement dangereuse pour l'humanité.

Sur la base de nos connaissances en neuroscience et en IA, il n'y a pas de raison de croire que l'intelligence humaine soit un sommet insurpassable d'intelligence, au contraire. Encore une fois, c'est excitant mais aussi inquiétant, si nous n'avançons pas sur le chemin de systèmes d'IA plus intelligents que nous avec la plus grande prudence. La théorie algorithmique en informatique nous suggère cependant que toute intelligence sera aussi limitée (et donc rend peu plausible le concept de singularité en IA), parce que nombreux problèmes computationnels sont irrémédiablement insolubles avec une puissance de calcul limitée : il y a donc aussi une limite supérieure à l'intelligence.

Jocelyn Maclure



De ses origines à aujourd'hui, la recherche en IA nous aide à mieux cerner les conditions de possibilité, les forces et les limites de l'intelligence animale, humaine et non humaine. Je comprends aisément que des chercheurs en informatique et des ingénieurs conjecturent que des programmes d'IA rendront éventuellement les machines approximativement aussi intelligentes que les humains (human-level, comme l'écrit Yoshua Bengio dans le texte d'introduction), ou peut-être même plus intelligentes (« superintelligence »). Il n'en demeure pas moins que l'intelligence animale est le fruit d'une longue évolution biologique ayant permis à certains organismes vivants de s'adapter à leur environnement et de se reproduire. Il est tout à fait plausible que certaines des propriétés que nous associons à l'intelligence ne soit pas réalisables sur la base d'une matière inorganique et en l'absence des cellules biologiques qui permettent à un organisme d'être en vie. Il est parfaitement possible d'être sceptique quant à l'atteinte de l'AGI tout en s'adossant à une conception entièrement naturaliste ou matérialiste du monde et de son fonctionnement.

La cognition humaine, en outre, n'implique pas seulement la capacité (réelle, mais faillible) de raisonner; elle implique aussi des désirs, des émotions, des intuitions. Elle est à la croisée de l'affectivité, de la conscience, de la volonté et de la raison; des notions à la fois essentielles à la compréhension de l'être humain et difficiles à cerner d'un point de vue scientifique. La cognition humaine est à la fois affaire de sens commun et de rationalité.

Les sciences cognitives ne sont plus exclusivement dominées par le behaviorisme et ses successeurs plus sophistiqués (fonctionnalisme, computationnalisme, etc.). La théorie triple E (embodied, embedded, enactive) de la cognition est depuis quelques décennies l'une des plus influentes. Cette théorie de la cognition inspirée de la phénoménologie allemande et française est sensible au rôle du corps (l'incarnation), de l'inscription dans un écolonnement à la fois naturel et culturel, et de l'(inter)action au sein de cet environnement. Il est envisageable que de donner un « corps » physique (robots) ou numérique (avatars) à une IA devant agir dans un environnement physique ou virtuel ne permette pas de combler le fossé ontologique entre l'artéfact et l'organisme vivant auquel j'ai fait allusion plus haut. Le fait que toutes ces théories demeurent pour le moment spéculatives ne les rend pas pour autant également vraisemblables à la lumière de ce que l'on connaît de la cognition humaine.

Q7

Comment pourrions-nous, en tant que communauté universitaire, mieux communiquer sur les questions relatives à l'IA afin d'éclairer le discours public?

Juliette Powell

En fin de compte, les gouvernements, les entreprises, les universitaires, les défenseuses et défenseurs de la justice sociale ainsi que les ingénieures et ingénieurs doivent non seulement établir une gouvernance interne et externe qui protège la majorité des personnes sur la planète, mais aussi communiquer de manière à ce que le grand public puisse comprendre. La communauté universitaire est parfaitement positionnée pour susciter un dialogue mondial qui permette à un plus grand nombre de personnes sur la planète de se poser de meilleures questions sur l'IA en tant que communauté mondiale. Par exemple, pourquoi développons-nous l'IA? Est-elle au service de l'espèce humaine ou de la course technologique? Construisons-nous l'IA pour la course aux armements ou pour l'espèce humaine?

Une autre question que doivent se poser tous les acteurs de l'IA, à tous les niveaux : Qu'est-ce que nous ne sommes PAS prêts à faire pour gagner de l'argent; qu'est-ce que nous ne sommes PAS prêts à faire pour obtenir un avantage concurrentiel; qu'est-ce que nous ne sommes PAS prêts à faire pour obtenir une suprématie mondiale en matière d'IA. À partir de là, la communauté universitaire interdisciplinaire doit communiquer les limites à respecter à la fois à l'interne et à l'externe, afin que nous puissions tous avoir le sentiment d'avoir notre mot à dire et d'être tenus responsables dans le cadre de ce nouveau contrat social mondial transparent. Comme le souligne Yosuha Bengio, une structure de gouvernance internationale garantirait que les droits de la personne soient au centre de toute gouvernance/réglementation de l'IA, qui serait activement appliquée en repérant et en restreignant les mauvais acteurs à tous les niveaux. Mais pour cela, il faut que les peuples du monde aient le sentiment de faire partie du processus sans avoir l'impression d'être laissés pour compte.

Hugo Loiseau

Les universitaires pourront mieux communiquer et éclairer non seulement les discours, mais aussi les décisions prises à propos de l'IA grâce à une méthode nommée Recherche et innovation responsables (RRI pour *Responsible Research and Innovation*) définie comme «un processus transparent et interactif par lequel les acteurs de la société et les innovateurs se répondent mutuellement en vue de l'acceptabilité (éthique), de la durabilité et de la désirabilité sociétale du processus d'innovation et de ses produits commercialisables (afin de permettre une intégration adéquate des avancées scientifiques et technologiques dans notre société)» (Von Schomberg, 2013).

Il faut donc travailler avec les outils théoriques et méthodologiques que nous avons à notre disposition à titre de chercheurs. Il faut travailler ce grand dialogue en interdisciplinarité et en intersectorialité (recherche, société, entreprises, gouvernements). La mise en œuvre est certes difficile et demandante, mais me semble possible dans un but ultime qui sera de développer la confiance entre tous ces acteurs à propos de l'IA. Cette méthode a déjà été utilisée avantageusement comme l'ont démontré le *UK Public Dialogue with Society*, les travaux de l'OBVIA et l'INGSA par exemple pour l'IA ou d'autres innovations technologiques.

Conclusion

Les réponses aux différentes questions ont permis d'offrir un regard transversal et interdisciplinaire sur les principaux risques soulevés par les récentes avancées en IA, mais également de réfléchir collectivement à la manière de prévenir ceux-ci et de s'assurer que le développement de cette technologie bénéficie à toutes et à tous. À cet égard, toutes et tous reconnaissent la nécessité d'une gouvernance internationale de l'IA, bien que l'importance de l'encadrement locale, régionale et nationale soit également mise de l'avant par les intervenants et intervenantes. Pour ce qui est des principaux enjeux identifiés, malgré quelques éléments de dissension, il a surtout été possible de voir émerger plusieurs points de convergence entre les réponses des différents intervenants et intervenantes, notamment en ce qui a trait aux enjeux touchant la démocratie et les droits humains.

Sommaires des principaux enjeux :

Démocratie et droit humains

- Usages malveillants, notamment de l'IA générative, menaçant le processus électoral et le débat public (production d'hypertrucages, création de faux comptes sur les médias sociaux)
- Utilisation de l'IA par les États comme instrument de domination et de contrôle, tant à l'interne (auprès de leur population) qu'à l'externe (cyberattaques entre États).
- Risques inhérents aux modèles d'IA comme les atteintes à la vie privée et les effets discriminatoires que produisent les algorithmes touchant principalement les personnes et les groupes les plus marginalisés et vulnérables.

À noter que plusieurs des risques menaçant la vie démocratique et les droits humains seraient par ailleurs exacerbés par la centralisation croissante du pouvoir, au sens où le développement de l'IA, bien qu'il affecte l'ensemble des individus, est (et sera de plus en plus) contrôlé par un petit nombre d'entreprises privées.

Coûts environnementaux et perte de contrôle des SIA

Au-delà des enjeux touchant la démocratie et les droits et libertés, les risques environnementaux découlant du coût énergétique des modèles d'IA ainsi que les dangers liés à une éventuelle perte de contrôle des systèmes d'IA, voire de « super intelligence » artificielle, sont également évoqués. Notons toutefois un point de désaccord en ce qui a trait à la plausibilité de cette perte de contrôle des systèmes d'IA, principalement au sujet de la priorisation des risques découlant de ce scénario. Cette divergence de point de vue semble s'expliquer, en partie du moins, par des visions différentes à propos des capacités de l'intelligence artificielle et de ce qui constitue l'intelligence humaine.

En terminant, nous remercions Yoshua Bengio, notre premier invité, ainsi que les intervenantes et intervenants d'avoir accepté de partager leurs réflexions. Ces échanges de perspectives et de points de vue différents, voire parfois opposés, ont permis d'éclairer le débat et d'enrichir notre compréhension des impacts sociétaux de l'IA et de la manière dont cette technologie transforme nos sociétés, et ce, dans un contexte de dialogue ouvert et respectueux. Cette initiative se veut ainsi le point de départ, la première étape vers la mise en place d'une collaboration accrue et d'une conversation continue entre spécialistes des sciences et génies, des sciences humaines et sociales et de la santé.

Références

Ada Lovelace Institute. (2023). *People, risk and the unique requirements of AI: 18 recommendations to strengthen the EU AI Act*. https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act/

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.

Browne, S. (2015). *Dark Matters: On the Surveillance of Blackness.* Durham: Duke University Press

Chun, W. H. K. (2015). On hypo-real models or global climate change: a challenge for the humanities. *Critical Inquiry*, 41(3), 675–703. https://doi.org/10.1086/680090

Chun, W. H. K. (2021). *Discriminating data: Correlation, neighborhoods, and the new politics of recognition*. MIT press.

Gebru, T. (2020). Race and Gender. Dans Dubber, M. D., Pasquale, F., et Das, S. (Eds.), *The Oxford Handbook of Ethics of AI* (Ser. Oxford handbooks). Oxford University Press.

Morin, E. (2017). *Science avec conscience*. Éditions Points. (ouvrage originale publié en 1982).

Noble, S. U. (2018*). Algorithms of oppression: how search engines reinforce racism*. New York University Press

Nye, J. S. (2011). Nuclear lessons for cyber security?. *Strategic studies quarterly*, 5(*4*), 18-38.

Quantum city (United Kingdom). (2020). *Public Dialogue Report on Quantum Technologies*. https://www.quantumcity.org.uk/news/public-dialogue-report-quantum-technologies

Von Schomberg, R. (2013). A Vision of responsible Innovation. Dans Richard Owen M. Heintz and J. Bessant (eds.), *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, pp. 51-74. Wiley.

Waltz, K. N. (1979). *Theory of international politics*. McGraw-Hill.

obvia IVADO Mila