« L'homme a perdu l'exclusivité de la parole qui héberge et véhicule le débat social »

Alexey Grinbaum, physicien et philosophe

Par **Nidal Taibi**

Encore un livre sur l'intelligence artificielle ? Depuis le « choc ChatGPT », il n'y en a plus que pour elle. Mais loin d'une redite, c'est une mise en perspective, civilisationnelle et anthropologique, que propose le physicien et philosophe Alexei Grinbaum. Discret médiatiquement, très prolifique dans le milieu académique et consulté en permanence par les organismes de l'éthique du numérique les plus réputés, l'auteur du tout récent *Parole de machines* (1) décrypte l'avènement d'une ère nouvelle, celle des machines parlantes qui, arrachant à l'homme le monopole de l'expression linguistique, bouleverseront notre langage et notre manière de penser. La frontière entre humain et non humain sera déplacée sans pour autant s'effacer, tient-il à rassurer.

Vous établissez d'entrée de jeu un parallèle entre le réchauffement climatique et le « réchauffement linguistique » provoqué par les machines parlantes. En quoi consiste ce parallèle ?

Qu'il s'agisse de la nature ou du langage, chacune de ces entités complexes évolue sous l'influence des technologies. Et ces évolutions, à la fois nous les subissons et les provoquons. Changement climatique : les êtres humains habitent une planète qu'ils modifient de plus en plus vite, au point de mettre leur habitat en péril. Changement linguistique : les êtres sociaux communiquent dans une langue qui accueille et véhicule notre histoire et notre culture et, cependant, les technologies dont nous sommes les concepteurs sont en train de supprimer notre monopole sur l'expression linguistique. Cela modifie et la nature, et le langage. Dans les deux cas, les conséquences ne sont pas minces : il s'agit d'une redéfinition de ce qu'est l'homme.

Comment l'IA et les machines parlantes provoquent-elles des changements anthropologiques et lesquels ?

La diffusion galopante de textes générés par les machines supprime le lien univoque entre le langage et l'être humain. Hannah Arendt a dit à la fin des années 1950 : « Tout ce que les hommes font, ou savent, ou ce dont ils ont l'expérience, a un sens seulement dans la mesure où il est possible d'en parler. » Cette donne millénaire est dorénavant obsolète. L'humain a perdu l'exclusivité de la parole qui héberge et véhicule le débat social. Mais l'objectif des ...

Le grand entretien

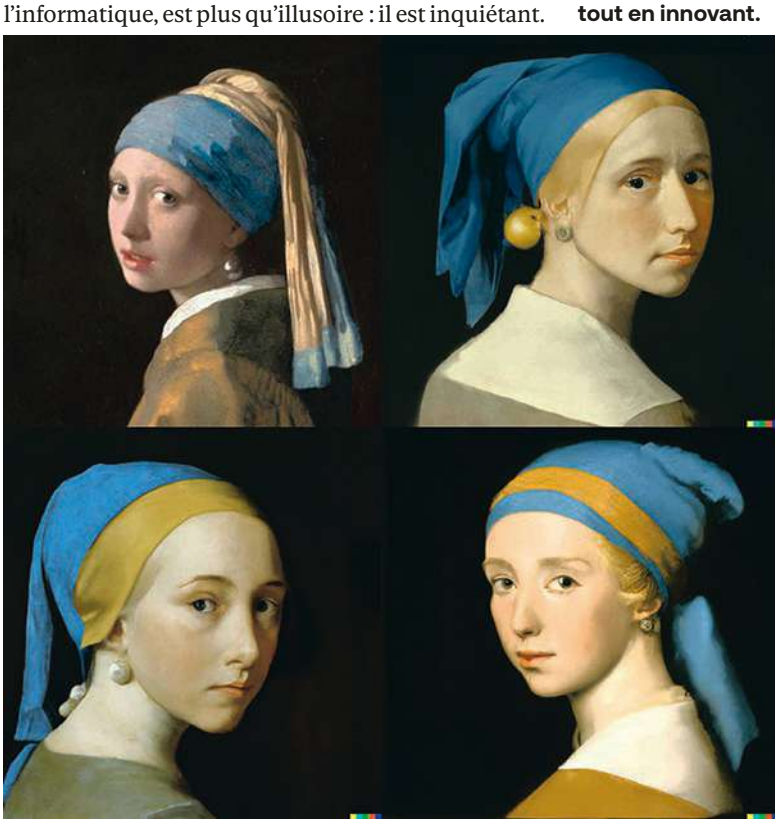
LAURENCE HONNORAT

… machines intelligentes n'est ni d'effacer notre civilisation ni d'oblitérer l'humain, comme le prophétisent certains. Les machines mettent en œuvre une finalité définie par le concepteur. Formulée dans un langage de programmation, cette finalité consiste à calculer et à optimiser des fonctions mathématiques et – sans surprise – ce processus purement formel produit des conséquences imprévues sur l'homme. Sur le plan socioéconomique, en modifiant plusieurs métiers, mais aussi sur le plan anthropologique, en accélérant la simplification de notre propre expression linguistique. La poésie disparaît de la langue. Et donc, les chemins de cette simplification sont tracés, non dans un monde dystopique où la pensée aurait cessé, comme se l'imaginaient les écrivains du XXe siècle, mais dans un univers du calcul froid, vide de toute signification humaine et étranger à l'éthique.

Cette simplification viserait selon vous à « aplatir la condition humaine en y soustrayant non seulement le mal mais aussi la ruse ou la séduction ». Vous dites que cela revient à « courir le danger d'un totalitarisme moral et émotionnel. » Qu'entendez-vous par là ?

Il faudrait cesser de vouloir faire croire aux bonnes gens que le monde virtuel peut être « éthique par conception ». Les systèmes d'IA ne connaissent pas le sens du bien et du mal. Pour gérer les conflits que provoquent les chatbots, on devrait emprunter un chemin radicalement différent de celui du système juridico-pénitentiaire dans le monde matériel. Ce désir de supprimer le mal de manière absolue, en corrigeant l'anthropologie humaine avec l'aide de l'informatique, est plus qu'illusoire : il est inquiétant.

L'IA ne répète pas bêtement. Elle est capable d'imiter tout en innovant.



DALL-E

Déjà, il mène à ce qu'on appelle en anglais *overpolicing*, ou excès de contrôle.

Dans quelle mesure ces machines parlantes modifient-elles la frontière entre humain et non humain ?

Les textes générés par la machine ressemblent parfaitement à ceux qu'aurait pu produire un humain. En lisant simplement un texte, il n'est désormais plus possible de saisir la différence. Par exemple, un professeur ne peut plus déterminer si une dissertation a été rédigée par un humain ou par une machine. Le besoin de maintenir ces distinctions pour les textes suffisamment longs est un impératif éthique. Il est nécessaire d'introduire des signes distinctifs dans les textes générés par les machines, sans pour autant nuire à l'utilité du chatbot. Techniquement, on ne peut jamais garantir à 100 % le maintien des distinctions, mais leur maintien, même imparfait, permet d'attribuer des responsabilités partagées en cas d'éventuel préjudice.

Ne s'agit-il pas toutefois de deux langages de natures différentes : d'un côté celui des machines, industrialisé, stéréotypé et utilitariste, de l'autre, le langage humain dont le propre est l'imprévisibilité et, éventuellement, la créativité ?

On a longtemps cru que les systèmes d'intelligence artificielle pourraient remplacer l'homme dans des tâches automatiques et routinières, tandis que les métiers créatifs resteraient à l'abri de l'automatisation. Aujourd'hui, nous savons que ce n'est pas le cas. L'IA générative est employée dans l'écriture d'articles de presse et dans le design de couvertures de magazines. La publicité se fait aujourd'hui avec, et parfois par, les créateurs artificiels. La machine est capable d'inventer quasiment sans limite, et sans que ses résultats soient trop éloignés de ce que l'utilisateur considère comme intéressant, compréhensible ou esthétiquement plaisant. Cette créativité extraordinaire, non humaine, n'a pas été programmée explicitement ; elle est le fruit inattendu d'un apprentissage suffisamment riche et complexe, qui donne à la machine la capacité, émergente et surprenante, de ne pas répéter bêtement les phrases de son corpus d'apprentissage mais de les imiter tout en innovant. Le résultat ressemble à une production humaine, même si le chemin que la machine emprunte pour y parvenir n'a rien à voir avec les méthodes d'apprentissage d'un cerveau. On ne fait que commencer d'en saisir les conséquences, y compris sur la poésie et sur notre créativité en tant que poètes de notre propre langue.

D'où « parlent » les machines parlantes de l'IA artificielle ? Y a-t-il, derrière, une idéologie implicite ? On les soupçonne de ne pas être neutres et de véhiculer l' « idéologie de la Silicon Valley »…

Les modèles d'intelligence artificielle parlent depuis un monde de pur calcul dans lequel il n'y a que des 0 et des 1. Ces machines ne possèdent ni connaissances, ni états d'âme, ni préférences politiques ou autres de l'humain. Mais l'utilisateur projette inévitablement sur

« L'objectif des machines intelligentes n'est ni d'effacer notre civilisation ni d'oblitérer l'homme. »

la machine une intériorité, une profondeur. Il l'accuse de biais ou de discrimination, et à raison : les sorties purement formelles d'un système d'IA acquièrent un sens humain aux yeux de celui qui les reçoit. Pour éviter des dommages trop importants, on ajoute aux modèles génératifs des filtres supplémentaires et des couches de contrôle éliminant tout langage jugé « toxique ». Mais c'est l'être humain qui décide ce qui est, ou non, toxique. Et donc, si les machines sont construites en Californie, les instructions pour éliminer le langage toxique portent inévitablement la marque des choix culturels californiens.

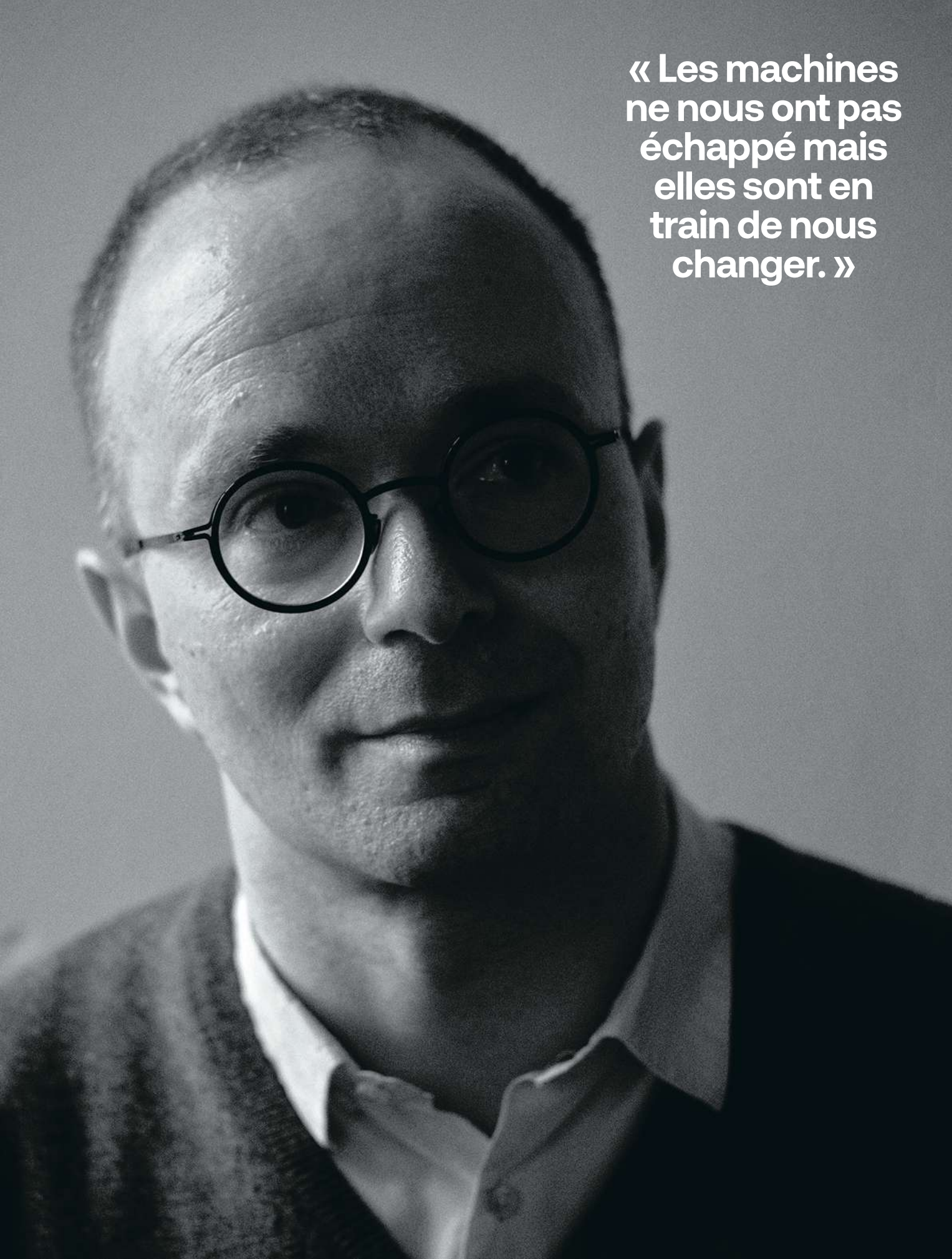
« La mutation du citoyen en utilisateur, de l'individu rationnel en individu numérique, change la société en profondeur », écrivez-vous. Quels changements sociaux et politiques induit cette mutation ?
L'utilisateur, par définition, cherche de l'utilité. En interagissant avec un système d'IA, il projette sur ce dernier des connaissances et aussi des émotions. Il est content quand « ça marche », quand la machine fait ce qu'elle est censée faire. L'utilisateur est quelqu'un d'affectueux, sensible, tantôt agacé, tantôt heureux, parfois emporté, souvent lassé. Rien de cela, à l'époque des Lumières, ne faisait partie des qualités requises pour être un bon citoyen ; le citoyen est un être éclairé, rationnel, informé, il est défenseur de valeurs et des libertés. Or, qui dit citoyen dit utilisateur. Les deux cohabitent dans un seul corps. Comment peuvent-ils cohabiter ainsi ? C'est la nouvelle tension anthropologique et politique de notre époque que doivent affronter tous les régimes, qu'ils soient démocratiques ou autoritaires.

Vous rappelez que l'humain n'a jamais eu le monopole de la parole. Des entités non humaines parlaient dans les mythes. Quelle particularité découle du fait que les machines deviennent parlantes ?
S'attendre à ce qu'une éthique de l'intelligence artificielle descende du ciel ou qu'on tombe dessus par hasard, c'est faire durer la paralysie de l'action devant le caractère inédit et stupéfiant de ces machines. La parole non humaine n'est clairement pas une nouveauté dans l'histoire de la pensée, et heureusement ! Autrefois, des entités non humaines qui peuplaient les mythes – dieux, anges ou démons – échangeaient avec les êtres humains par la bouche des oracles ou dans les songes. Ils s'exprimaient dans la même langue. Aujourd'hui, ce rôle revient aux machines. Pour que les nouveaux récits technologiques aient pour nous un sens, il est impératif de les mettre en lien avec les récits anciens qui fondent notre histoire en tant que civilisation et humanité.

Justement, les références théologiques et bibliques parsèment votre réflexion. Dans quelle mesure peuvent-elles nourrir la réflexion sur les enjeux contemporains autour de l'IA et des machines parlantes ?
Les récits technologiques sont tout aussi humains, et tout aussi inhumains, que les mythes qui parlent de dieux ou de démons. On a même inventé pour cela un terme technique : « comportement émergent ». Ce sont des capacités non explicitement programmées qui apparaissent à partir de l'interaction complexe entre les opérations de calcul élémentaires. Dans les machines, il n'y a que des 0 et des 1, et pourtant elles peuvent nous mentir ou produire des résultats remplis de beauté. De la même manière, notre monde est fait de la matière, d'atomes et de particules élémentaires, mais à un « niveau émergent » nous parlons, à travers et dans le langage, de dieux, anges ou démons. Comme le dit un texte ancien, « nous n'usons pas de simples mots, mais de sons tout remplis d'efficace ». Avec les machines, nous voyons que notre langage est aussi efficace, il provoque une action de la part de la machine, même si un système d'intelligence artificielle n'y voit aucune signification.

« Il nous faut trouver le chemin vers un* Nouveau Testament *de l'intelligence artificielle. » Qu'entendez-vous par là ?
Des lamentations, nous en entendons depuis Mathusalem. Il y a toujours eu des visions utopiques et dystopiques. Mais que les machines risquent de nous remplacer, cela reste une fiction. Leur fonctionnement est clairement différent de la manière humaine de manier le langage. La génération d'un texte n'est pas une affaire de sémantique et la machine ne connaît pas le sens littéral des mots. C'est l'utilisateur qui projette spontanément un sens sur le langage. Ces illusions provoquées par les chatbots – il s'agit bien d'illusions – recèlent néanmoins un pouvoir sur l'homme et la société. Certaines sont heureuses ou bienfaisantes, d'autres se révèlent manipulatrices ou maléfiques. N'importe quel système d'intelligence artificielle, même fabriqué dans les meilleures intentions, pourrait mentir, et cela n'est pas un scandale. Nous allons nous constituer de nouveaux récits qui permettront de donner un sens à ce monde nouveau où nous entrons avec les machines parlantes.

En filigrane de votre réflexion, on retrouve l'idée que les machines vont trop vite pour les humains. Est-ce encore possible de les maîtriser, de les encadrer, ou la créature a-t-elle déjà échappé à son maître ?
Adin Steinsaltz, remarquable érudit talmudique, rappelle une interprétation peu connue des textes sacrés. Elle est sans doute ancienne, parce que le motif qu'elle utilise est commun à l'islam et au judaïsme. Cette interprétation consiste en une phrase : la faute **...**


« Les machines ne nous ont pas échappé mais elles sont en train de nous changer. »

... réaction trop rapide, viscérale, de l'utilisateur humain qui ne prend pas assez de temps pour réfléchir. Mais la vitesse inhumaine du calcul est, dans le même temps, source de l'efficacité des machines et de leur utilité. C'est un vrai paradoxe : nous ne pouvons pas aller à la même vitesse or ce décalage entre deux vitesses, celle de notre action technologique sur le monde et celle de notre capacité d'en prévoir les conséquences, apparaît comme source de tous les problèmes éthiques. Les machines ne nous ont pas échappé mais elles sont en train de nous changer, et cela va très vite, peut-être trop.

Que vous inspirent les récents appels et initiatives, dont la pétition initiée par Elon Musk, qui appellent à un moratoire sur la recherche sur l'IA ?
Cet appel au moratoire sur la construction de modèles de langages « plus puissants que le GPT-4 », n'est pas un exemple unique dans l'histoire de la technologie. Il s'inscrit plutôt dans une lignée d'appels similaires à des moratoires sur la recherche dans le domaine des nouvelles technologies, notamment en ingénierie génétique ou dans la modification du génome humain. Comme d'autres technologies à double usage, l'IA générative nécessite une gouvernance internationale et un ensemble de mesures réglementaires. Il est certain qu'un cadre réglementaire, quel qu'il soit, contiendra de nombreuses lacunes : les effets à long terme échappent souvent au contrôle par la loi et tout cadre réglementaire dans le contexte géopolitique de la course à l'IA n'aura que des conséquences pratiques limitées. Toutefois, malgré ces limites, un cadre pour l'IA générative servira deux objectifs essentiels et nous avons besoin d'un peu de temps pour le dessiner. Premièrement, il reflétera la prise de conscience politique du fait que l'IA générative joue un rôle majeur dans la vie de la société. Sa valeur symbolique ne doit pas être sous-estimée, même si la recherche scientifique et l'innovation technologique ambitieuse doivent continuer. Deuxièmement, un cadre réglementaire établira les critères pour l'attribution des responsabilités en cas de dysfonctionnement ou de dommages provoqués par l'utilisation de l'IA.

Vous appelez à une éthique du numérique. Quels en seraient les principes ?
Il existe de nombreuses listes de principes éthiques pour l'intelligence artificielle. Littéralement, plusieurs centaines ! La transparence des systèmes d'IA, leur explicabilité, etc. Mais l'éthique du numérique ne se réduit pas à une liste des valeurs ou des principes. Il y a des conflits ou des tensions entre plusieurs valeurs, et elles jouent un rôle essentiel. L'éthique du numérique est avant tout une affaire de réflexion, et non celle d'un jugement de conformité. Je ne vois pas d'autre solution que de ménager aux machines une place au sein des récits qui ont contribué à former le socle de notre culture et de notre civilisation. Les concepts moraux n'existent pas dans la nature. On les trouve dans les récits que les hommes se racontent depuis la nuit des temps et qui introduisent des motifs fondamentaux, ceux-là mêmes que l'on distingue aussi dans notre réalité technologique.

D'un point de vue politique et juridique, comment peut-on encadrer les machines parlantes, et l'IA en général ?
A l'échelle européenne, le travail est en cours pour établir une nouvelle loi sur l'IA. La Commission a fait une proposition en avril 2021, puis le Conseil européen en décembre 2022. C'est maintenant au tour du Parlement de se prononcer sur le texte. Celui-ci inclut plusieurs innovations : une hiérarchie des systèmes d'IA en fonction des risques qu'ils posent, des « regulatory sandbox » *(NDLR : des espaces, conçus et contrôlés par un organisme de réglementation, destinés à mettre à l'essai de nouveaux processus sous la supervision de cet organisme avant leur entrée sur le marché)* pour soutenir l'innovation et des articles très contraignants sur l'IA générative à la suite de la parution de ChatGPT. Quel sera l'impact de cette loi ? Il est trop tôt pour le prédire, mais l'Europe essaie clairement de trouver son propre chemin en la matière et de se démarquer des exemples de la réglementation chinoise ou américaine.

Quels sont les enjeux et défis, éthiques et politiques, auxquels devrait face la réglementation européenne ?
On peut comparer la nouvelle réglementation chinoise de l'IA générative avec la loi européenne qui sera soumise au vote cette année. Les Chinois sont allés très vite en publiant un ensemble de 21 articles qui font peser une responsabilité inouïe sur les fournisseurs de ces systèmes et exigent qu'ils mettent en œuvre un grand nombre de contrôles dont certains techniquement difficiles, voire impossibles, à garantir. Le brouillon actuel de la loi européenne dit la même chose. Même si le fournisseur ne peut aucunement prévoir telle ou telle sortie individuelle et qu'il n'avait pas l'intention de provoquer un dommage, il en porte néanmoins la responsabilité. Cette responsabilisation totale du fournisseur fait sens sur le plan juridique, mais elle va beaucoup étonner les ingénieurs qui ne se perçoivent pas comme coupables puisque le comportement d'un système d'IA comme ChatGPT reste – même pour eux ! – complètement imprévisible. En revanche, la différence essentielle entre les cas chinois et européen se trouve au plan politique plutôt que technique. Le texte chinois évoque les « valeurs socialistes fondamentales » et protège explicitement le pouvoir de l'Etat. Le brouillon européen met l'accent sur la protection des droits de l'homme. D'ailleurs, au Parlement européen, ce texte est suivi non seulement par la commission du Marché intérieur et de la protection des consommateurs mais aussi par la commission des Libertés civiles. Mais l'accord est difficile à trouver parce que l'idée de réglementer par une loi un modèle fondamental de l'IA, et non un produit final mis surle marché, est assez inédite dans l'histoire du droit. ●

(1) *Parole de machines*, par Alexei Grinbaum, humenSciences, 192 p.

Bio express

1978
Naissance à Leningrad, en Russie, le 30 novembre.

2004
Obtient un doctorat en « économie et sciences sociales » à l'Ecole polytechnique.

2006
Devient chercheur au CEA (Commissariat à l'énergie atomique).

2016
Est nommé expert auprès de la Commission européenne.

2019
Intègre le Comité national pilote d'éthique du numérique.

2022
Devient président du Comité opérationnel d'éthique du numérique du CEA.

GLEB KALININ