L'IA peut-elle être **responsable ?**

GÉNÉRAL D'ARMÉE (2S) WATIN-AUGOUARD
Fondateur du Forum InCyber

MAKÉDA PECASTAING
Directrice de la communication du Forum InCyber

JUIN 2024

agora
INCYBER FORUM

INTRO-DUCTION

L'intelligence artificielle subjugue en même temps qu'elle inquiète. Son immense potentiel, n'est pas encore pleinement exploré dans sa dimension, sa puissance, son impact.

Elle nous transporte dans l'inconnu, dans un futur imprévisible, laissant la voie ouverte à toutes les hypothèses, au discours des Cassandre ou des utopistes.

L'intelligence, dont la tentative de définition souligne la complexité, est sans aucun doute ce qui distingue l'humain de toute autre espèce vivante, même s'il existe une intelligence animale.

Avec l'intelligence artificielle, désincarnée, nous extériorisons ce qui prend naissance dans la profondeur de notre cerveau. Avec elle s'opère une dissociation du corps et de la pensée, illustrée notamment par le robot autonome, doté d'une intelligence que certains imaginent égale voire supérieure à la nôtre et prenant forme humanoïde. « Robots animés, avez-vous donc une âme qui s'attache à notre âme et la force de nous surpasser ? », écrirait aujourd'hui le poète Alfred de Musset.

Ainsi pourrait s'accomplir, si l'on n'y prenait garde, un « grand remplacement » de l'humain par la machine, le dépassement du point de singularité en quelque sorte, la fin du corps « enveloppe » de l'esprit, « l'euthanasie de la mort » pour reprendre l'expression de Laurent Alexandre.

Toutes les technologies, dans leur phase d'émergence, ont suscité des interrogations, des peurs, fragilisant parfois leur développement initial. N'a-t-on pas craint que les vaches ne produisent plus de lait à cause du chemin de fer ? Plus récemment, le Covid-19 n'a-t-il pas été causé par certains à cause de la 5G ? Entre l'obscurantisme qui leur attribue tous les malheurs du monde et la naïveté que nourrit l'émerveillement face à un « Nouveau Monde », le point d'équilibre doit être trouvé.

L'intelligence artificielle connaît cette phase de doute et d'exaltation. Le discours s'enrichit chaque jour de prédictions qui mêlent raison et fantasme.

Depuis que l'intelligence artificielle est sortie de son « hiver », depuis que les connexionnistes ont développé les systèmes neuronaux, autorisant l'apprentissage profond, depuis l'émergence et surtout, de sa diffusion dans le grand public de l'intelligence artificielle générative, préalable, selon certains à une intelligence artificielle générale, le spectre des prises de position couvre tout et son contraire.

Une chose est certaine, plus que toute autre technologie, l'intelligence artificielle impose que l'on porte sur elle toute notre intelligence humaine, en perçant le nuage de l'invisible et du mystérieux. Plus que toute autre technologie, elle invite à répondre à la question « pourquoi ? », c'est-à-dire une recherche de sa finalité, d'un encadrement qui préserve notre liberté, sans jamais porter atteinte à ce qui fait l'essence même de l'humain : son identité – c'est-à-dire son unicité – son intimité, sa liberté de penser, de rêver, de s'exprimer, de décider. Ajoutons sa conscience, mot qui vient du latin *scire* (« connaître »), et nous renvoie à notre compréhension du monde, à notre intuition. C'est le sens de la célèbre pensée de Rabelais : « Science sans conscience n'est que ruine de l'âme ! »

Cette interrogation quasi existentielle appelle le développement d'une IA responsable, encadrée, mais non limitée dans sa capacité innovatrice, par un droit inspiré par l'éthique. À dire vrai peut-on faire nôtre l'expression « IA responsable » qui laisserait accroire qu'une IA peut être responsable au même titre qu'une personne physique ou morale ? C'est la thèse soutenue par certains, prônant la création en droit d'une « personnalité robotique » à côté de la dualité personnalité physique/personnalité morale. Un certain anthropomorphisme se dégage souvent des termes et expressions utilisées : non seulement, on attribue à l'IA une capacité spécifiquement humaine mais également une apparence humaine avec les robots.

Cette approche a été partagée un moment par le Parlement européen, dans un rapport de 2017 ; le règlement européen sur l'IA l'a heureusement écartée. L'IA n'est pas responsable, puisque nous le sommes lors de toutes les phases, depuis sa conception jusqu'à sa mise en œuvre.

Le développement de l'intelligence artificielle ne peut être envisagé sans une cybersécurité, garante de la confiance. La cybersécurité doit être au service de l'IA. L'IA doit être au service de la cybersécurité. Les deux entretiennent un rapport adversaire/partenaire. Les systèmes d'IA doivent être sûrs et sécurisés, tout au long de leur cycle de vie afin d'être résistants face aux cyberattaques. Celles-ci peuvent atteindre les modèles de fondation, la « boîte noire », les résultats.

L'intelligence artificielle offre de nombreuses opportunités aux cybercriminels pour identifier et exploiter des vulnérabilités, déjouer les systèmes de détection, agir sur un très grand nombre de cibles, identifier des voies d'attaque.

Les menaces identifiées comprennent, entre autres, les reprogrammations, l'empoisonnement des données, leur exfiltration et toute autre infection, les attaques antagonistes qui trompent ou détournent un modèle d'apprentissage automatique, les pirates tentant d'inverser les modèles d'IA pour rechercher les données d'entrainement.

La preuve est aujourd'hui faite également que certaines IA peuvent apporter aux prédateurs des solutions prêtes à l'emploi pour construire des cyberattaques (*phishing*, injection de *malware*, etc.). *Deepfakes* (hypertrucages) et *deepvoices* sont également utilisés pour tromper sur l'identité et permettre des cyberattaques par altération de la confiance.

Mais l'IA est aussi une alliée précieuse de la cybersécurité. Plus nous nous dirigeons vers un monde hyperconnecté, plus les systèmes cyberphysiques seront développés, plus l'IA nous aidera à avoir une réponse systémique offrant des capacités de résilience des systèmes complexes, tels les espaces dits intelligents.

Détection d'anomalies, de menaces, analyse comportementale, analyse prédictive, approche globale automatisée, aide à la décision sont des voies de progrès déjà intégrées par les solutions de renseignement sur les menaces cyber, les EDR, l'analyse des voies d'attaque (*Attack Path Management*).

L'intelligence artificielle rebat les cartes entre la capacité offensive des cybercriminels et la capacité de défense des acteurs. Entre l'attaquant et le défenseur s'instaure le duel « du canon et de la cuirasse », une course aux armements qui est asymétrique, car le prédateur n'a guère de scrupules au regard des contraintes juridiques et éthiques que nous nous fixons.

Avec l'IA, nous ne nous inscrivons pas dans un parcours linéaire mais dans une dynamique exponentielle. Jamais sans doute, depuis les origines du Forum, nous n'avons été placés devant autant d'enjeux, géopolitiques, industriels, sociétaux, sécuritaires. Nous ne sommes pas démunis comme le montre le génie inventif des entreprises présentes au sein du Forum InCyber, comme le soulignent les interventions des experts venant de 103 pays, lors de l'édition 2024. Une chose est certaine : l'intelligence humaine devra toujours dominer l'intelligence artificielle. Pour cela, il nous faut démystifier, apprendre pour comprendre et sans doute plus que jamais replacer l'humain au cœur de notre cybersécurité. Cette posture déterminée et optimiste se nourrit de l'esprit du Forum qui transcende la forme. Oui ! Le Forum InCyber c'est d'abord un état d'esprit, une quête de l'intérêt général, une mission de service public qui associe tous les acteurs publics et privés, civils et militaires.

La 16e édition qui s'est tenue à Lille, du 26 au 28 mars 2024, a bien sûr porté sur les aspects techniques de l'IA. Mais elle a aussi été l'occasion d'échanger en toute liberté lors de l'Agora qui, autour de parlementaires et d'acteurs de la société civile, a fait salle comble. L'atelier PhilosoFIC, traditionnel espace de débat, a complété la réflexion collective. C'est pourquoi cette publication de l'Agora InCyber l'associe pour enrichir son contenu.

GÉNÉRAL D'ARMÉE (2S) WATIN-AUGOUARD
Fondateur du Forum InCyber

MAKÉDA PECASTAING
Directrice de la communication du Forum InCyber

Le panel des intervenants

AGORA

Mounir Belhamiti, Député de la Loire-Atlantique

Mireille Clapot, Députée de la Drôme

Catherine Morin-Desailly, Sénatrice de la Seine-Maritime

Miguel Ángel Cañada, Head of National Coordination Centre (NCC-ES) at INCIBE

Jérôme Clauzade, Chief Product Officer, Crowdsec

Jean-Philippe Desbiolles et **Grégoire Colombet,** coauteurs de *Humain ou IA ? Qui décidera le futur ?*, Éditions Dunod. Prix du livre Forum InCyber « Grand Public »

Emmanuelle Legrand, Magistrate, ancienne négociatrice IA Act (FR)

Thiébaut Meyer, Director, Office of the CISO, Google Cloud

Sawsen Rezig, CTO & Co-founder ShareID

Eric Salobir, Président du comité exécutif de la Human Technology Foundation, membre du CNNum

Michel Séjean, Professeur des universités, spécialiste de droit de la cybersécurité et de droit comparé. Directeur scientifique du *Code de la cybersécurité* (Dalloz)

Emmanuel Saliot, Conseiller sécurité et nouvelles technologies au Parlement européen

Rayna Stamboliyska, Fondatrice de RS Strategy

Benoît Tabaka, Secrétaire général, Google France

PHILOSOFIC

Emilie Bonnefoy, CEO d'Open Sesam

Cécile Doutriaux, avocate au barreau de Strasbourg

Guy-Philippe Goldstein

Sandrine Hilaire, responsable de la commission « Confiance numérique d'e-Futura

Tariq Krim

Jean Peeters, ancien président de l'université de Bretagne sud (UBS) titulaire de la chaire *Souveraineté numérique et cybersécurité* de l'IHEDN

SOMMAIRE

L'INTELLIGENCE ARTIFICIELLE, **UNE TECHNOLOGIE DISRUPTIVE ?**

« Intelligence artificielle » n'est pas une formule qui s'inscrit dans la tendance, comme la plupart des mod Si elle est au cœur de tous les débats, c'est notamme parce que chacun prend conscience d'une lame de fon de nature à « submerger » notre société, à mettre ses paradigmes à l'épreuve. Cette prise de conscience était réservée aux spécialistes, avant que la démocratisation de l'IA générative, fin novembre 2022, ne fasse éclater aux yeux de tous la puissance, mais aussi les risques de l'IA.

L'IA est-elle une technologie de rupture qui va changer notre vie, nos rapports au travail, nos rapports à la formation, à l'information, aux autres ?

Une innovation déjà **ancienne**

Nous semblons découvrir l'intelligence artificielle depuis la diffusion, en quelques jours, de ChatGPT, une IA générative créée par OpenAI, association fondée par Elon Musk en 2015. Le 30 novembre 2022, ChatGPT nous dit : « Je suis un modèle de langage informatique conçu par OpenAI. Mon but est de pouvoir répondre aux questions et fournir de l'aide aux utilisateurs en exploitant mon apprentissage automatique et mes connaissances en langage naturel. J'essaie de fournir des réponses précises et utiles aux utilisateurs. Je suis un outil de traitement de langage naturel et je n'ai pas d'opinions personnelles ni de préférences ». Il lui a fallu cinq jours pour compter 1 million d'utilisateurs ; Facebook avait dû attendre 10 mois pour le même résultat…

L'historien est toujours en quête de l'origine : à quel moment le raisonnement humain s'inscrit-il dans une démarche intellectuelle qui est la semence de l'IA ? Les travaux de Pascal, de Leibniz, de Joseph Marie Jacquard, de Charles Babbage assisté sur sa « machine analytique » par Ada Lovelace, préfigurent l'idée de confier à la machine des tâches que l'humain accomplit lentement ou péniblement. Mais le véritable commencement est sans doute à rechercher dans l'article publié par Alan Turing, en 1950[1], dans lequel il posait la question suivante : « Les machines peuvent-elles penser ? ». Il répondait en ces termes : « Nous pouvons espérer que les machines pourront égaler les hommes dans tous les domaines purement intellectuels[2]. »

Mais le commencement date sans aucun doute de la conférence au Dartmouth College (É.-U.), réunissant du 11 au 16 août 1956, plusieurs chercheurs, notamment Marvin Minsky, Claude Shannon, Nathan Rochester, Allen Newell et John McCarthy. Cette conférence pose les fondations d'une nouvelle discipline de recherche et utilise pour la première fois l'expression « intelligence artificielle », inventée par McCarthy. Pour ce dernier, relève de l'intelligence artificielle une « machine qui soit douée de langage, puisse former des concepts et des abstractions, et soit capable de résoudre des problèmes que pour l'instant l'homme est seul à pouvoir traiter ». L'intelligence artificielle est, selon lui, « la construction de programmes informatiques qui s'adonnent à des tâches qui sont pour l'instant accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique ». Pour McCarthy, l'intelligence artificielle doit « procéder sur la base de la conjecture que chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut en principe être décrit avec une précision telle qu'une machine peut être faite pour le simuler ». Le terme « simuler » est important, car il montre que les pionniers imaginaient une limite par rapport à ceux qui évoquent parfois la « conscience de l'IA ».

Depuis cette conférence, deux écoles se développent. La première est incarnée par le courant symboliste qui domine de 1956 à 1990. Selon Daniel Andler[3], « l'IA symbolique cherche à comprendre et à reproduire la dynamique des pensées en prenant pour unité élémentaire des propositions simples qui sont ensuite manipulées selon des règles formelles. »

1 Alan Mathison Turing, « Computing Machinery and Intelligence », Mind, 59, octobre 1950, p. 433-460.

2 Alan TURING, *Computing Machinery and Intelligence,* MIND, 1950.

3 Daniel ANDLER, *Intelligence artificielle, intelligence humaine,* NRF Essais, Gallimard, 2023.

L'IA symbolique a recours à des connaissances, des règles logiques et des symboles (concepts, objets, relations). Elle est principalement développée dans les « systèmes experts ». Elle répond aux sollicitations par le biais d'un moteur d'inférence, logiciel correspondant à un algorithme de simulation des raisonnements déductifs, qui applique les règles logiques de la base aux connaissances déjà inscrites dans la base et à la nouvelle question qui lui est soumise. Elle peine cependant à traiter les données non structurées (images, son, etc.)

La deuxième école est portée par le courant « connexionniste », inspiré par l'article écrit par Warren McCulloch, spécialiste des neurosciences, et Walter Pitts, spécialiste de la logique, en décembre 1943[4]. Ces chercheurs imaginaient un système cognitif sous forme d'un réseau de neurones artificiels, semblable à celui du cerveau humain. En 1957, Frank Rosenblatt, au sein du laboratoire aéronautique de Cornell, met au point le *Perceptron*, premier système d'apprentissage reposant sur la connexion de neurones artificiels. L'école symbolique et l'école connexionniste se sont développées au même moment, mais les connexionnistes n'ont guère suscité d'enthousiasme, quand ils n'ont pas été critiqués par les symbolistes.

L'intérêt initial, suscité par l'intelligence artificielle, s'est atténué. Dix ans après Dartmouth, le National Research Concil's Automatic Language Porcessing Advisory Comittee publie un rapport négatif qui entraine des coupes budgétaires dans le secteur. L'ARPA, l'agence américaine du Pentagone, financée par le budget militaire, refuse de financer les projets non-militaires, dont ceux liés à l'IA. Mais c'est surtout une intervention sur la BBC, le 30 août 1973, du mathématicien anglais, sir James Lighthill, qui sonne le glas. Celui-ci a publié un rapport[5] pour le British Science Research Council sur l'évaluation de la recherche universitaire dans le domaine de l'intelligence artificielle.

Il y écrit : « Les étudiants [en IA] concluent souvent qu'il est irréaliste d'espérer le développement d'ici la fin du XXe siècle d'un système généralisé capable de manipuler une grande base de connaissances. » Cette vision très pessimiste conduit alors le gouvernement britannique à mettre un terme au soutien de la recherche sur l'IA dans les universités.

Un second « hiver de l'IA » a lieu à la fin des années 1980 (1987-1993). La relance l'IA dans les années 1990 est due en particulier à la robotique et l'apprentissage profond (*deep learning*) qui bénéficie de la profusion de données et de la puissance du calcul de haute performance (HPC). Les symbolistes, qui ont dominé jusqu'alors, voient la revanche des connexionnistes. Pour autant, ils n'ont pas totalement perdu la partie puisque l'intelligence artificielle symbolique s'insère dans les réseaux de neurones profonds.

La mise sur le marché de ChatGPT, fin novembre 2022, s'inscrit dans la continuité mais est annonciatrice de grands changements, ne serait-ce qu'en raison de la prise de conscience par tous d'enjeux qui ne faisaient débat qu'au sein de groupes d'experts. Elle montre, en moins d'un an, au travers des évolutions, que l'IA générative n'est pas le terme d'un processus mais le début d'une métamorphose de l'IA, sous l'impulsion de nouveaux usages, notamment grand public, qui stimulent le développement des technologies. L'IA peut aussi, selon certains, migrer vers une intelligence générale. Celle-ci pourrait égaler, voire surpasser l'intelligence humaine. Mais cette étape, incertaine pour beaucoup de scientifiques, n'est pas encore atteinte. Toutes les mesures prises aujourd'hui pour réguler l'IA ont notamment pour objectif de contrer le mythe de Prométhée.

4 "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics.*

5 *"Artificial Intelligence : A General Survey".*

Un essai de **définition de l'IA**

Le règlement européen *AI Act* définit la notion de système d'intelligence artificielle (« SIA ») comme étant tout « système basé sur une machine, conçu pour fonctionner avec différents niveaux d'autonomie et qui peut faire preuve d'adaptabilité après son déploiement et qui, pour des objectifs explicites ou implicites, déduit, à partir des informations qu'il reçoit : comment générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions pouvant influencer les environnements physiques ou virtuels ». Si nous prenons la définition communément admise, l'IA est « un ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine ». En procédant presque à du copié-collé, l'Union européenne s'est alignée sur la définition de l'OCDE pour qui : « Un système d'intelligence artificielle est un système automatisé qui, pour des objectifs explicites ou implicites, déduit, à partir d'entrées reçues, comment générer des résultats en sortie tels que des prévisions, des contenus, des recommandations ou des décisions qui peuvent influer sur des environnements physiques ou virtuels. Différents systèmes d'IA présentent des degrés variables d'autonomie et d'adaptabilité après déploiement ».

Comme le souligne le général Perrot, conseiller IA du Commandement cyber du ministère de l'Intérieur (ComcyberMi), le terme générique d'IA dissimule de nombreuses disciplines, de nombreuses méthodes comme de nombreux champs applicatifs. « L'IA doit être perçue comme un amplificateur de l'intelligence humaine développé à partir de méthodes mathématiques auto-apprenantes. Elle est donc un agglomérat de méthodes mathématiques qui conviennent plus ou moins à la modélisation du problème posé. La question sous-jacente à l'utilisation de l'IA est donc d'abord mathématique. Il s'agit de savoir comment approximer des données non linéaires dans un espace de grande dimension. Pour accomplir ce « miracle », l'IA fait appel à de nombreuses disciplines : l'analyse harmonique, la morphologie, les statistiques, les probabilités, la théorie de la décision, la géométrie, les systèmes dynamiques… De toutes ces méthodes, la plus pertinente pour éclairer le mystère mathématique qui entoure l'IA est certainement à rechercher du côté de la géométrie par la caractérisation de formes invariantes dans des données communes. Mais les mathématiques à ce jour ne sont pas parvenues à résoudre le problème de la non linéarité en grande dimension. C'est du côté de l'observation et donc de la physique qu'il a fallu s'orienter pour trouver une solution : l'apprentissage automatique et notamment via sa dernière déclinaison l'apprentissage profond ».

Daniel Andler[6] conforte cette pluridisciplinarité : « C'est une erreur de penser que l'intelligence artificielle est quelque chose d'homogène et de clairement défini ».

Cédric Villani[7] considère que l'intelligence artificielle désigne en effet moins un champ de recherches bien défini qu'un programme, fondé autour d'un objectif ambitieux : comprendre comment fonctionne la cognition humaine et la repro-

6 Daniel ANDLER, *Intelligence artificielle, intelligence humaine,* op. cit.

7 Cédric VILLANI, *Donner un sens à intelligence artificielle, rapport,* 2018.

duire ; créer des processus cognitifs comparables à ceux de l'être humain. La pensée connexionniste est ici bien exprimée. Le champ de l'IA est donc naturellement extrêmement vaste, tant en ce qui concerne les procédures techniques utilisées que les disciplines convoquées : mathématiques, informatiques, sciences cognitives… Les méthodes d'IA sont très nombreuses et diverses (ontologique, apprentissage par renforcement, apprentissage adversarial, réseaux de neurones, etc.). « Ce n'est pas – selon le mathématicien – une technique secrète qui a été mise au point par tel ou tel laboratoire, ce n'est pas un secret jalousement gardé, c'est un ensemble de techniques très variées, avec des convergences, avec des stratégies différentes, qui vous permettent d'obtenir un très haut niveau d'efficacité dans certaines tâches algorithmiques, et qui permettent de faire des tâches qu'on aurait crues réservées aux humains, naïvement ».

Nous le voyons, il faut déjà beaucoup d'intelligence humaine pour comprendre ce qu'est l'intelligence artificielle et de mesurer pleinement la révolution anthropologique structurelle que celle-ci engendre.

Une **révolution anthropologique** structurelle

Selon **Jean-Philippe Desbiolles**[8], l'IA crée une rupture structurelle et non conjoncturelle. « C'est la fin du code ; on passe d'un monde déterministe qui a régi les 80 dernières années à un monde d'apprentissage. C'est une révolution !

Nous n'avions jamais bâti des systèmes capables d'apprendre au fur et à mesure de leur utilisation. Nous nous basions sur des systèmes déterministes à base de règles qui s'appuyaient sur la relation « *if/ then* ». Ce monde-là va cohabiter avec un monde d'apprentissage. Alors que l'on parle beaucoup de technologie, de *data* et de *data scientist*, etc., le sujet majeur, pour moi, c'est nous, c'est l'humain. Pourquoi ? Parce que qui dit apprentissage, dit pédagogie. Qui dit pédagogie, dit phénomène d'apprentissage. Qui apprend à l'IA? L'humain, nous. Donc nous sommes les tenants et aboutissants de cette révolution ». Il poursuit : « L'IA générative n'est pas une rupture technologique. Celle-ci n'est pas encore arrivée parce que l'IA générative est toujours un système probabiliste, ni plus ni moins. Mais préparons-nous. La prochaine révolution sera beaucoup plus violente, de mon point de vue, mais chaque chose en son temps. Rien n'a changé, mais tout a changé. La révolution de la génération actuelle est une révolution des usages, non de la technologie ».

La révolution de la génération actuelle est une révolution des usages, non de la technologie.

Jean-Philippe Desbiolles

8 Jean-Philippe DESBIOLLES est coauteur de *Humain ou IA, qui décidera le futur ?*, Éditions DUNOD, 2023.

Le terme « révolution » est également employé par **Eric Salobir** qui n'exclut pas la technologie de son champ : « Nous vivons certes une révolution technologique, une révolution économique. Mais pour moi, c'est avant tout une révolution anthropologique. C'est une transformation en profondeur de notre relation au monde. On ne s'en rend pas compte parce que c'est le début. Cela peut sembler être un peu pompeux de dire les choses ainsi. En vérité, je pense qu'il y a une peur d'un grand remplacement par la machine. Cela vient de la façon dont on anthropomorphise la machine, que l'on perçoit un peu comme une espèce de compétiteur. Ce compétiteur devient le compétiteur des cols blancs, ce qui n'est plus indifférent. Tant que les cols bleus étaient les seuls concernés, certains étaient un petit peu plus à l'aise. Mais aujourd'hui, beaucoup ont peur d'être eux-mêmes concernés par une disruption. Pour beaucoup de nos interlocuteurs et de nos contemporains, c'est aussi une machine qui parlerait avec un côté un peu "magique", mais je pense qu'on ne s'est pas complètement départi d'une forme de pensée magique, quasiment «totémique», avec, de ce fait, une confiance faite à la machine par beaucoup de nos contemporains, confiance probablement un petit peu exagérée, Donc une révolution avec des systèmes qui sont extrêmement performants, mais qui balbutient encore et qui sont encore très loin de ce qu'ils feront dans quelques années ». **Eric Salobir** nous invite donc d'éviter d'avoir un rapport un peu « fétichiste » à ce type de technologie et de la regarder pour ce qu'elle est vraiment.

Nous vivons le dépassement de la finitude : gains de productivité infinis, dépassement des ressources, prolongation de la vie jusqu'à l'immortalité ?

Emmanuel Saliot

Emmanuel Saliot constate aussi des effets transformationnels très puissants sur deux domaines en particulier. Tout d'abord, souligne-t-il, « nous vivons le dépassement de la finitude : gains de productivité infinis, dépassement des ressources, prolongation de la vie jusqu'à l'immortalité ? Avec l'IA, on entretient un peu ce mythe de l'expansionnisme qui appelle un regard sur la consommation et sur les ressources. Deuxième domaine assez dangereux : la taylorisation de l'émotion. Celle-ci entretient un rapport émotionnel avec la machine et crée un brouillard sur la vérité, avec une difficulté aujourd'hui à savoir non plus ce qui est vrai, mais qui est vrai ».

L'IA crée une relation inédite entre l'humain et la machine. La machine était conçue pour servir l'humain. L'IA pourrait-elle l'asservir ?

L'IA crée une relation inédite entre l'humain et la machine. La machine était conçue pour servir l'humain. L'IA pourrait-elle l'asservir ?

Humain et IA : adversaires partenaires ?

Dès l'origine, la relation entre l'humain et la machine est au cœur du développement de l'IA, censée remplacer le premier pour accomplir des tâches pénibles, répétitives. C'est l'ambition d'Alan Turing et des pionniers de Dartmouth. Mais avec l'IA générative, la question se pose en d'autres termes, avec une autre intensité. L'IA passe-t-elle du statut d'assistant à celui de remplaçant ? Élargit-elle son champ d'action du monde physique, dont la robotisation est un exemple, vers le domaine de la pensée ?

DÉCIDER, PENSER, AGIR À LA PLACE DES HUMAINS ?

Le débat public est souvent assez binaire, souligne **Emmanuelle Legrand**, avec des approches antinomiques qui opposent ceux qui, à propos de la relation humain-machine, pensent que « l'IA c'est bien » et ceux qui pensent que « c'est mal ». « Nous dérivons parfois très vite sur le fait de savoir effectivement si la machine est seule, autonome, mais il ne faut pas confondre l'entraî-

nement et le fait que la machine ne s'auto-crée pas toute seule. Il y a un humain qui crée une machine, et c'est bien une machine ».

Aussi longtemps que l'intelligence artificielle demeure un outil qui aide l'humain à automatiser des tâches intellectuelles, c'est un bénéfice pour la société, selon **Mireille Clapot** qui considère que la rupture vient lorsque la technologie commence à prendre des décisions à notre place. « Tant qu'elle aide l'humain dans l'accomplissement de tâches fastidieuses ou pénibles, le problème ne se pose pas. Mais ce n'est plus le cas lorsque l'on constate que, dans certains pays, des jugements sont rendus par des machines, par des IA.

Par exemple, un robot juge est disponible 24 heures sur 24 pour traiter des litiges civils en Chine par des machines, par des IA et non par des humains ».

Le remplacement de l'humain pour des tâches monotones et régulières n'est peut-être pas une conséquence inéluctable. Certaines entreprises maintiennent des tâches industrielles qui pourraient être robotisées pour que certains puissent s'accomplir dignement dans le travail[9].

Catherine Morin Desailly souligne l'impact binaire de l'IA : « Je crois qu'on ne mesure pas encore les conséquences que l'IA va avoir sur nos vies, sur nos modèles politiques, nos modèles de société, notre modèle civilisationnel et le devenir de l'homme aussi, tout simplement. Notre devoir, c'est de se demander quelle sera la place de l'homme dans ce monde. Est-ce que l'IA servira le progrès ? Je crois qu'il y a des choses formidables devant nous, prenons la médecine, par exemple, mais il y a des menaces et des risques qu'il faut savoir anticiper ». Il faut donc replacer l'humain au cœur de l'IA et surtout le faire de manière collective, car son développement est un problème de société.

L'intelligence artificielle touche à ce que nous avons de plus unique : le langage et, en amont, la réflexion autour de la perception que nous avons du monde ?

Emilie Bonnefoy

Citant Voltaire, Émillie Bonnefoy rappelle que « Peu de gens s'avisent d'avoir une notion bien entendue de ce qu'est l'humain. Cet extrait de *Doutes sur l'Homme* (1er chapitre de son *Traité de métaphysique*) montre que, déjà à cette époque, on s'interrogeait sur ce qui fait sa spécificité, son caractère unique. L'intelligence artificielle touche à ce que nous avons de plus unique : le langage et, en amont, la réflexion autour de la perception que nous avons du monde. L'intelligence humaine se manifeste à travers deux spécificités : la capacité à penser et la capacité de perception. Cette capacité à penser et à percevoir le monde s'exprime ensuite à travers le langage qui permet de "mettre en musique sa relation au monde", pour reprendre l'expression d'Asma Mellah ».

Cette réflexion autour d'une intelligence artificielle qui pourrait « supplanter » l'intelligence humaine, faire mieux qu'elle, produire du langage à sa place, nous renvoie à notre propre condition.

Jean Peeters partage la même perspective lorsqu'il évoque l'impact de l'IA sur la socialité. « L'artificialisation crée un langage et ce langage agit sur la socialité, c'est-à-dire sur la tendance des individus à vivre en groupe ; elle formate le raisonnement, crée un mode de pensée. Le langage est la réflexion de la réalité mais il en est aussi le constructeur. Celui de l'intelligence artificielle s'adresse aux individus, aux groupes, aux sociétés ; il reconstruit la réalité d'une façon qui pose question ».

9 Jean-Michel OUGHOURLIAN, *Le travail qui guérit l'individu, l'entreprise, la société,* Éditions PLON, 2022.

Jean-Philippe Desbiolles apporte une contribution nuancée sur le rapport entre l'humain et la machine : « L'IA générative porte bien son nom : elle génère. Elle génère du contenu au sens large, quel que soit le contenu, quel que soit le format, etc. Jusqu'à présent, on connaissait l'IA dite "traditionnelle", sur le modèle du *machine learning* qui travaillait sur des contenus qui étaient produits par nous, les humains. Ce paradigme est en train de nous transporter dans une autre réalité – il a peut-être déjà basculé – parce que l'IA générative génère du contenu, qui sert à entraîner d'autres systèmes et à leur apprendre à fonctionner. Donc les machines apprennent à des machines. D'où la notion de la compliance et de la responsabilité qu'il faut maintenant mettre en regard avec cet état de fait. C'est bien ou mal, mais c'est ainsi. Nous vivons dans une ère où les machines produisent du contenu qui sert à entraîner les machines.

À la vitesse à laquelle elles se développent, poursuit-il, ces machines ne s'arrêtent pas, elles ne dorment pas, elles ne fatiguent pas et elles continuent à générer. Donc, mécaniquement, nous sommes dépassés ». Cela fait douze ans que **Jean-Philippe Desbiolles** travaille sur l'IA ; les sept premières années, il se disait assez serein sur la notion de complémentarité humain-IA puisque la performance de l'humain plus l'IA était mécaniquement, statistiquement, plus forte, plus élevée que l'IA seule ou l'humain seul. Pour lui, ce paradigme n'est plus vrai aujourd'hui. « Aujourd'hui, dans certains cas d'application, l'IA seule a une performance bien plus élevée que l'humain. Dans d'autres cas, l'humain seul a une performance bien plus élevée que l'IA. Enfin, nous avons des cas d'application, largement majoritaires aujourd'hui, où la complémentarité entre l'humain et l'IA a une performance supérieure à la seule IA ou au seul humain. Il s'agit donc dans chaque cas d'application de retenir le dispositif le plus efficient, le plus optimal et de vérifier, du point de vue sociétal ou légal, si nous le tolérons l'acceptons ou le rejetons ».

Il faut préserver notre libre arbitre dans la manière dont on fait collaborer des utilisateurs avec ces systèmes algorithmiques.

Grégoire Colombet

Selon **Grégoire Colombet**[10] : « Nous pouvons prendre autant de décisions que l'on veut à un moment donné avec un système algorithmique ; à la fin il y aura toujours un humain qui est responsable. Quand on parle de conscience, il faut avoir les moyens de cette conscience parce que, finalement, on s'est aperçu que la collaboration entre l'humain et la machine n'est pas évidente. Il faut préserver notre libre arbitre dans la manière dont on fait collaborer des utilisateurs avec ces systèmes algorithmiques. Parce qu'on est influencé, parce qu'on se laisse entraîner par la machine.

Et donc c'est tout l'enjeu derrière cette révolution de l'intelligence artificielle générative avec laquelle il est difficile parfois d'exercer son esprit critique parce que le contenu est complexe, parce que le contenu est long, etc ».

Emmanuel Saliot montre aussi son inquiétude. Il pense qu'il sera de plus en plus difficile de maintenir l'humain dans la boucle : « Quand vous regardez le nombre de paramètres aujourd'hui dans les modèles d'IA génératifs, c'est un peu comme le "chiliogone" de Descartes[11]. Nous n'arriverons pas à avoir un contrôle vraiment humain. S'agissant d'expérience sur les binômages d'IA, ceux-ci sont meilleurs pour améliorer les prompts, très largement devant les humains.

10 Grégoire Colombet est coauteur de *Humain ou IA, qui décidera le futur ?*, Éditions DUNOD, 2023.

11 Dans ses *Méditations métaphysiques*, Descartes montre à partir d'un chiliogone (polygone à 1000 côtés) que des choses faciles à concevoir sont parfois difficiles à représenter.

Donc, la communication machine-machine est déjà plus performante. La vitesse de décision prise sur le champ de bataille sera telle, je le pense, que les hommes n'arriveront pas à suivre le rythme. *Human-in-the-loop*[12], la grande sauvegarde, va être difficile, voire une illusion. C'est vrai que l'IA ne pense pas ; elle est une réduction du monde statistique, mais on a observé ce qu'on appelle les capacités émergentes, des capacités qu'on ne comprend pas vraiment. Notamment quand elle traduit d'une langue à une autre, elle passe par des représentations qu'on ne comprend pas ; elle est capable d'inventer des langages, mais on ne sait pas très bien comment. Enfin, il faut regarder aussi à l'étranger : au Brésil, par exemple, le Parlement étudie un projet de loi sur la propriété intellectuelle des innovations créées par les IA qui possèderaient elle-même le brevet[13] ».

Laisser l'IA décider à la place des humains soulève de nombreuses questions en temps « normal ». Dans les situations paroxystiques, comme les conflits, cette substitution remet en cause les principes fondamentaux de la guerre. **Mounir Belhamiti**, membre de la Commission de la défense et des forces armées à l'Assemblée nationale, apporte un regard inquiet sur les applications de l'IA sur le champ de bataille : « Je vois passer des analyses prospectives d'usage d'IA dans le secteur de la défense qui sont pilotées par des hommes et des femmes avec des finalités bien précises qui inquiètent. Je vous le dis clairement : au-delà de remettre l'humain au cœur de l'IA, il y a nécessité à encadrer l'usage des IA dans certaines applications, et notamment les applications militaires. Avec l'IA, avec les IA, nous allons observer demain un changement significatif sur le champ de bataille et plus largement sur la scène géopolitique internationale. C'est cela qui m'inquiète[14] ». La crainte du député se rapporte notamment à la question des systèmes d'armes létales autonomes (SALA) qui peuvent tuer sans intervention humaine. La guerre est sans aucun doute l'activité humaine la plus inhumaine. Si l'IA s'y invite…

Tariq Krim partage ce regard pessimiste : « Maintenant, nous sommes confrontés à des drones autonomes, sur des zones où un être humain aura une durée de vie de quelques minutes puisqu'il est pris en compte par des drones qui analysent, frappent ; ce sont des drones autonomes. C'est le cadre horrible des guerres qui sont en train de se dérouler. L'Ukraine est assurément un point de bascule ».

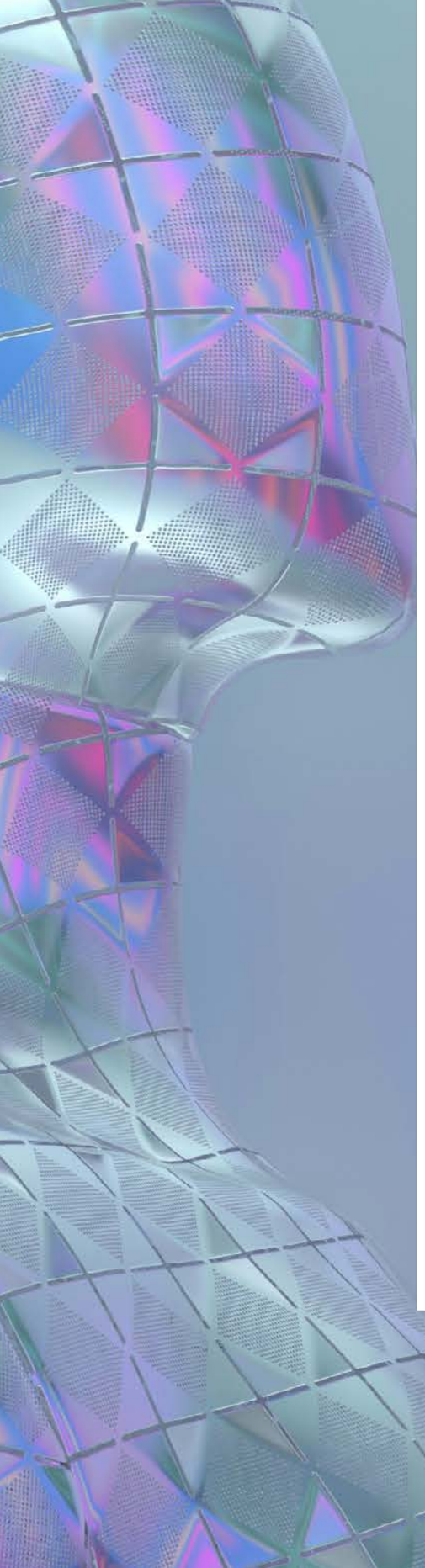
12 *Human-in-the-loop* (HITL), modèle d'apprentissage automatique dans lequel l'intelligence humaine est intégrée au processus de décision, permettant une interaction humain-machine.

13 Dans une décision d'un tribunal de Virginie, rendue jeudi 2 septembre 2021, la juge fait notamment valoir que la loi américaine requiert qu'un individu prête serment quand il fait sa demande de brevet, et qu'un individu est par définition une personne humaine. « Au fur et à mesure que la technologie évolue, il peut arriver un moment où l'intelligence artificielle atteindra un niveau de sophistication qui pourrait satisfaire les significations acceptées de la qualité d'inventeur », a déclaré la magistrate dans la décision.

14 La crainte exprimée par Mounir Belhamiti est illustrée par le programme israélien Lavender qui automatise en large partie le ciblage et a donc permis le « traitement d'un nombre beaucoup plus important de cibles en peu de temps ». (https://www.972mag.com/lavender-ai-israeli-army-gaza/)

Avec les IA, nous allons observer demain un changeur significatif sur le champ de bataille et plus largement sur la scène géopolitique internationale.

Mounir Belhamiti



UNE MENACE POUR LA DÉMOCRATIE ?

Si l'IA peut encadrer, diriger, modeler la pensée au travers notamment du conditionnement algorithmique, il est un domaine particulièrement sensible celui de la démocratie qui s'exprime en particulier à l'occasion des échéances électorales. Mais c'est aussi la remise en cause des élites, en général, et des élus en particulier qui peut susciter une crainte exprimée par des parlementaires.

Mireille Clapot s'inquiète des effets de l'IA sur la démocratie « Nous, qui sommes des politiques, nous voyons arriver des intelligences artificielles qui se targuent de faire mieux que nous, en se prétendant très fiables, parce qu' "insensibles à la corruption". Là, nous atteignons les fondements même de la démocratie et tous les pans de la société. Donc, c'est là qu'il faut être très vigilants ». Cette inquiétude est partagée par **Catherine Morin-Desailly** : « Ce qui me préoccupe à l'heure actuelle, c'est la menace visant notre démocratie : explosion des fausses nouvelles, de *deepfakes*, manipulation des discours, faux discours, fausses prises de parole, etc. Comment peut-on peut lutter contre tous ces phénomènes ? Nos démocraties ne sont-elles pas gravement en danger ? »

La démocratie repose, en effet, sur le libre arbitre, sur l'autonomie de décision des citoyens au moment des élections. Son fonctionnement est tributaire de la confiance dans les représentants de la nation. Ce regard pessimiste de deux parlementaires met en exergue la question de l'autonomie de la pensée à l'ère de l'IA.

Quel avenir pour la **pensée ?**

L'intelligence artificielle générative se développe à partir d'immenses bases de données (LLM). Les sources sont multiples (bases de données, contenus en ligne, réseaux sociaux, etc.) ; elles sont analysées par le *deep learning* qui les contextualise. Si les données d'apprentissage sont mauvaises, si elles sont appauvries lorsqu'elles proviennent de contenus issus d'autres IA, les modèles de langage produiront une réponse erronée.

L'IA s'appuie sur des données, traitées par des algorithmes et sur lesquelles s'appliquent des calculs statistiques. Sans données, il n'y a pas d'IA, ce qui explique en partie « l'hiver de l'IA », à l'époque où la création de données était encore trop faible. En 2030, on estime la production annuelle à 1000 milliards de téraoctets. « Nous sommes dans un monde numérique, de plus en plus, de part en part. Un monde de données. Ces données qui sont au cœur du fonctionnement des intelligences artificielles actuelles », selon Cédric Villani.

L'IA ET NOS DONNÉES À CARACTÈRE PERSONNEL

Toutes les données n'ont pas la même valeur. Certaines font l'objet d'une protection particulière. Il en est ainsi des données à caractère personnel qui relèvent du RGPD et ne peuvent faire l'objet d'un traitement sans le consentement des personnes, chaque fois que le règlement l'impose. Les modèles de fondation de l'IA peuvent contenir de nombreuses indications permettant d'identifier directement ou indirectement des personnes au sens du règlement.

Ainsi, en mai 2022, l'Information Commissioner's Office (ICO) a ordonné à la société américaine Clearview AI de supprimer les données des résidents britanniques de ses systèmes de reconnaissance faciale. Cette entreprise a collecté, sans leur consentement, plus de 20 milliards d'images de visages de personnes sur Facebook, sur les réseaux sociaux et sur le Web. Le RGPD impose par ailleurs un principe d'exactitude des données selon lequel les données traitées doivent être exactes et tenues à jour. La Commission nationale informatique et libertés (CNIL), dans une note du 11 octobre 2023, considère que le RGPD offre un cadre innovant et protecteur pour l'IA.

LE **RGPD**, CADRE INNOVANT ET PROTECTEUR POUR L'IA

SOURCE : NOTE DE LA CNIL DU 11 OCTOBRE 2023

Le principe de finalité impose de n'utiliser des données personnelles que pour un objectif précis (finalité) défini à l'avance. La CNIL admet qu'un opérateur ne puisse pas définir au stade de l'entraînement de l'algorithme l'ensemble de ses applications futures, à condition que le type de système et les principales fonctionnalités envisageables aient été bien définies.

Le principe de minimisation n'empêche pas l'entraînement d'algorithmes sur de très grands ensembles de données. Les données utilisées devront en revanche, en principe, avoir été sélectionnées pour optimiser l'entraînement de l'algorithme tout en évitant l'utilisation de données personnelles inutiles. Dans tous les cas, certaines précautions pour assurer la sécurité des données sont indispensables.

Le principe de conservation limitée n'empêchera pas la définition de durées longues pour les bases de données d'entraînement, qui requièrent un investissement scientifique et financier important et deviennent parfois des standards largement utilisés par la communauté.

La réutilisation de bases de données est possible dans de nombreux cas. La réutilisation de jeux de données, notamment de données publiquement accessibles sur Internet, est possible pour entraîner des IA, sous réserve de vérifier que les données n'ont pas été collectées de manière manifestement illicites et que la finalité de réutilisation est compatible avec la collecte initiale.

La profusion de la *data*, liée avec le calcul de haute performance (HPC) rend possible des créations qui ne sortent pas du néant mais puisent dans toutes les ressources déjà créées, stockées, échangées. Ces contenus sont, avec ce que nous avons acquis par nous-mêmes (formation, expérience, etc.), le terreau de notre pensée. En nous offrant des solutions « sur étagère », l'IA peut encourager notre paresse et donc réduire les fonctions intellectuelles qui modèlent notre intelligence. Elle est un prédateur pour la création littéraire et artistique, car les modèles de fondation vont puiser dans l'existant, sans véritablement s'interroger sur les droits des auteurs. S'agissant des contenus, ils peuvent être le fruit de données volontairement modifiées ou créées à des fins malveillantes pour altérer le résultat. Enfin, des biais peuvent aussi contribuer à une information décalée.

DONNÉES PILLÉES, MODIFIÉES, BIAISÉES

Les données de référence sont parfois pillées. Dès le lancement des IA génératives, des contentieux sont apparus au regard de la propriété intellectuelle. La création d'un *Next Rembrandt* à partir de plus de 160 000 fragments issus de 346 toiles du maître détenues par des musées porte-elle atteinte aux droits de ces derniers ?

Le *New York Times* a porté plainte devant la cour fédérale du district de Manhattan contre OpenAI et Microsoft pour avoir utilisé ses articles par millions pour entraîner leur famille de grands modèles de langage utilisée par ChatGPT, Bing Chat et Copilot. Le journal estime que ces entreprises n'ont pas respecté le copyright protégeant ses archives et ont ainsi créé les conditions d'une concurrence déloyale. « Bien que les accusés se soient livrés à des copies à grande échelle à partir de nombreuses sources, ils ont accordé une importance particulière au contenu du *Times* lors de l'élaboration de leurs LLM, révélant ainsi une préférence qui reconnaît la valeur de ces œuvres ». OpenAI et Microsoft, son principal investisseur, sont aussi visés par une nouvelle plainte aux États-Unis, déposée par des publications[15] pour atteinte au droit d'auteur. Une proposition de loi française propose de compléter l'article L. 131-3 du code de la propriété intellectuelle par un alinéa ainsi rédigé : « L'intégration par un logiciel d'intelligence artificielle d'œuvres de l'esprit protégées par le droit d'auteur dans son système et a fortiori leur exploitation est soumise aux dispositions générales du présent code et donc à autorisation des auteurs ou ayants droits ». L'*AI Act* oblige les concepteurs à identifier les documents protégés par le droit d'auteur.

Comment déterminer dans un agglomérat de contenu ce qui relève de tel ou tel auteur, de tel article de presse, etc. ?

Cécile Doutriaux

15 *Chicago Tribune, New York Daily News, Denver Post, Orlando Sentinel, Sun Sentinel of Florida, San Jose Mercury News, Orange County Register et St. Paul Pioneer Press.*

Cécile Doutriaux estime qu'il y a un véritable problème pour ceux qui ont écrit des articles, composé des partitions, peint ou dessiné, c'est-à-dire les journalistes, les écrivains, les auteurs.

Comment déterminer dans un agglomérat de contenu ce qui relève de tel ou tel auteur, de tel article de presse, etc. ?

Toute la culture du monde peut être phagocytée, digérée.

Tariq Krim

Tariq Krim est plus accusateur : « Les plateformes ont pris la culture mondiale, l'ensemble des choses qui constituent notre culture, que ce soient des photos, des images, des vidéos, etc., on a scanné tout YouTube, tout Wikipédia. La plupart des plateformes ont pompé tout le contenu en piratant et ont fait la même chose avec toute la littérature mondiale. Cela s'appelle *Common Crawl*. Toute la culture du monde peut être phagocytée, digérée ». On pourrait citer à propos d'emprunts, la voix de Scarlett Johansson qu'OpenAI aurait imitée pour son assistant personnel.

Il y a donc un équilibre à trouver. **Benoît Tabaka** confirme la nécessité de donner des possibilités de contrôle aux titulaires de droit, aux créateurs, afin qu'ils puissent continuer à créer et à vivre de leur création. « Mais la mise en place de murs peut conduire à l'accès non seulement du contenu mais aussi du savoir qui est inclus dans ce contenu. Être capable de dissocier le savoir du contenu, c'est-à-dire de l'information versus l'élément protégé par la propriété intellectuelle, va être un des enjeux de demain ».

Eric Salobir est-il réaliste ou pessimiste lorsqu'il « se rend compte qu'on est face à une espèce de paradoxe, parce que d'un côté, nous cherchons à protéger les droits d'auteur, ce qui est tout à fait légitime, et, de l'autre, nous observons que des pays comme le Japon disent entraîner les IA sur autant de données que l'on souhaite, gratuitement. Ainsi, le moment venu, il y aura quelque chose de japonais dans les modèles. Le danger serait de construire une espèce de Tour de Babel qui soit finalement très anglophone, qui soit très nord-américaine ». Il appuie cette observation sur ses souvenirs : « J'ai travaillé avec un artiste nord-africain qui était résident à la Villa Médicis, à Rome. Il me disait que lorsqu'il essayait de créer des univers arabes à partir d'une IA, celle-ci lui proposait automatiquement Tatooine, c'est-à-dire la planète des sables dans *Star Wars*. Donc l'IA concevait américain. La question est de savoir comment on réussit à inclure notre culture dans ces grands modèles et, en même temps, à protéger les intérêts légitimes de nos artistes ».

Cela prouve bien que les IA génératives actuelles sont de la pure statistique et vont accroître le déterminisme des décisions qui seront basées dessus. Sur le plan géopolitique, ces IA vont donc accroître la domination des grands empires, notamment américains et chinois.

Les données peuvent être modifiées pour tromper, causer un préjudice. **Cécile Doutriaux** l'évoque au travers de l'hypertrucage : prendre l'image d'une personne, faire un montage que l'on anime. Il y a une usurpation d'identité, une atteinte au droit à l'image, à, la dignité humaine. Mais le système peut lui-même être à l'origine de pertes d'informations.

Tariq Krim illustre la déperdition du savoir en prenant l'image du griot, le vieux sage des villages de l'Afrique de l'Ouest qui raconte l'histoire du village : « Il s'en rappelle parce que son grand-père lui en a parlé, et puis il répète les choses. De temps en temps, l'histoire change un peu. Il l'embellit. Il n'aime pas telle personne, donc elle n'existe plus. D'une certaine manière, on va se trouver dans une situation assez fascinante.

Les *Large Language Models* sont l'équivalent de ce vieux sage. La question, c'est qu'on ne sait plus si ce qu'il dit est vrai ou faux ». Il poursuit : « Préserve-t-on la culture sous sa forme originale, avec des archives, des bibliothèques ? Ou comme dans *Fahrenheit 451*, considère-t-on qu'elles n'ont plus de nécessité ? L'IA nous explique des choses. On ne sait pas si c'est vrai ou faux dès lors que nous n'avons plus accès à la source initiale ».

De la falsification, de la perte aux biais, il y a peu de distance, dans la mesure où ces trois déviations contribuent à l'altération de la vérité.

Cécile Doutriaux craint l'appauvrissement de la pensée humaine, parce que les données d'entraînement sont souvent puisées dans des contenus du passé. « La capacité humaine, l'intelligence, c'est aussi la faculté d'anticiper. Ce n'est pas d'être dans un monde probabiliste. Certains disent que l'intelligence artificielle n'est pas que générative, elle peut être créatrice. Oui, mais que sur des bases du passé ».

Benoît Tabaka soulève aussi le problème de l'antériorité des données : « Lorsqu'on entraîne des grands modèles de langage uniquement sur des œuvres données, élevées dans le domaine public, le risque est grand d'avoir une réponse sur le statut de la femme, l'esclavage, la peine de mort, très différente par rapport à la réalité du jour, puisqu'on pourrait avoir 60-70 ans de retard par rapport à l'actualité ».

Sandrine Hilaire rappelle qu'Amazon avait conçu une IA pour le traitement des CV. L'entreprise ne recevait que des CV masculins car l'IA avait été entraînée sur les CV de ses développeurs qui étaient presque tous des hommes.

VERS LA PENSÉE UNIQUE ?

Tariq Krim qualifie l'IA de « grossesse statistique » : « On mixe des contenus pour en sortir la tendance majoritaire. Pour que l'humain s'oppose à la tendance majoritaire, il lui faut beaucoup de personnalité. Notre autonomie cognitive est altérée parce qu'une machine a appris pour nous et nous payons pour qu'elle le fasse. Est-ce de la productivité ou est-ce de la paresse ? Nous construisons des systèmes qui vont comprendre comment nous fonctionnons et, une fois qu'ils ont compris, nous devenons prisonniers, qu'on le veuille ou non. Je ne peux m'empêcher de *scroller*[16], car ils ont été conçus pour être addictifs ».

Sandrine Hilaire estime que, plus que jamais, la question du libre arbitre est essentielle. « Quand des étudiants utilisent ChatGPT, on constate que petit à petit tout le monde fait les mêmes réponses. Le plus important, ce qui fait notre force, c'est ce qu'il y a à l'intérieur de nous et ce que nous pouvons en faire ».

Emilie Bonnefoy évoque un enjeu civilisationnel. « Je pense en particulier à Socrate, qui incarne la culture du doute : la capacité qu'a l'être humain de construire un raisonnement, de s'appuyer sur un raisonnement didactique, de questionner le monde qui l'entoure et d'expliquer un raisonnement. Ce côté "boîte noire" doit nous rappeler que fondamentalement, notre capacité d'être humain, c'est celle de remettre en question les choses, de douter. « Tout ce que je sais, c'est que je ne sais rien », disait Socrate. Je pense qu'il faut revenir à un fondamental qui est la question de l'explicabilité des décisions et, en particulier, à la phase de démonstration que nous avons tous pratiquée en mathématiques ».

Notre autonomie cognitive est altérée parce qu'une machine a appris pour nous et nous payons pour qu'elle le fasse.

16 *Scroller* : anglicisme. Faire défiler un contenu sur un écran informatique.

Cécile Doutriaux craint un discours aseptisé : « Où est la fantaisie ? Où est le sens de la dérision ? Où est l'humour ? Il y a quand même des choses irréductibles chez nous. Et on dit qu'il faut créer la confiance. On dit que la confiance est l'une des possibilités divines de l'homme. C'était de Montherlant. Alors soyons divins, on verra ce que cela donne ». **Cécile Doutriaux** porte ainsi le discours sur le terrain de la philosophie.

Elle souligne ainsi que l'IA ressort aussi des sciences cognitives, des sciences humaines et sociales qui tentent d'expliquer ce qu'est l'humain dans son unicité, sa singularité. Pour préserver nos cerveaux contre une aseptisation de la pensée, il convient de les protéger, de les durcir. Le cerveau est le système de traitement automatisé de données le plus sophistiqué mais aussi le moins bien protégé face aux attaques[17]. Il faut donc le défendre par l'éducation et la formation.

Pour avoir conscience de ce qui est en train de se produire, il faut comprendre. Et pour comprendre, il faut être formé, éduqué.

Catherine Morin-Desailly

L'ÉDUCATION ET LA FORMATION PROTECTRICES DU LIBRE ARBITRE

Catherine Morin-Desailly cite Rabelais, l'homme de la Renaissance, selon lequel, en cette grande période de l'humanisme, « science sans conscience n'est que ruine de l'âme ». « Sans conscience nous renvoie à la morale et à l'éthique, mais l'expression signifie aussi sans mesurer, sans analyser, sans anticiper ce qui va se passer. Donc, notre devoir c'est de dire quelle sera la place de l'homme dans ce monde. Est-ce que l'IA servira le progrès ? » Elle poursuit : « Pour avoir conscience de ce qui est en train de se produire, il faut comprendre. Et pour comprendre, il faut être formé, éduqué. Le risque, c'est le décrochage, d'abord entre les pays riches. Sur l'IA, les États-Unis et la Chine se livrent un combat sans merci. L'Europe arrive à tirer, bon an mal an, son épingle du jeu. C'est un vrai défi, nous devons tout de même le dire. Mais beaucoup de pays à travers le monde sont en voie de décrochage. Le risque est grand d'accroître la pauvreté, les possibilités de développement et de solidarité à travers la planète. Le monde de demain risque d'être gouverné par une élite technologique qui comprendra tout, tandis que les autres seront dans l'ignorance.

Quand nous agissons dans le domaine politique, nous devons faire des efforts très importants pour comprendre ce qui est en train de nous arriver. Certains discours, lorsque je les entends, éveillent en moi quelque chose dont je n'avais pas du tout mesuré l'impact. Cette nécessité de monter en compétences numériques par tous, c'est vraiment un impératif, non seulement à l'école, pour la formation initiale ou continue, mais aussi en dehors, pour au moins prendre ensemble les bonnes décisions et maîtriser notre destin numérique. Ne laissons pas les décisions prises aux mains d'une élite technologique ».

17 Marc Watin-Augouard, « Alerte sur le système de traitement automatisé le plus sophistiqué et le moins bien protégé », InCyber News.

L'acceptation du public est liée à la transparence à son égard. De la même façon qu'on ne peut priver du droit à la parole un non spécialiste du nucléaire, une personne qui n'est pas spécialiste de l'intelligence artificielle doit être considérée.

Mireille Clapot

Mireille Clapot partage cette opinion tout en appelant de ses vœux une montée en compétence des femmes, encore trop minoritaires parmi les acteurs du numérique. Pour la députée, « il faut aussi tenir l'ensemble de la chaîne de la formation en étant vigilants sur le risque d'affaissement des capacités cognitives, notamment de nos jeunes. En particulier, en prenant l'exemple de la calculatrice, il faut continuer à apprendre le calcul mental afin de savoir encore compter et vérifier que les ordres de grandeur sont les bons, si un jour, nous sommes privés de calculette, Il faut aussi développer l'esprit critique chez nos jeunes – c'est un peu le pendant du calcul mental – pour pouvoir contrôler ce que sort une IA. Il faut aussi penser à tout ce public qui est éloigné du numérique. Nous devons regarder le milieu de la chaîne. L'acceptation du public est liée à la transparence à son égard. De la même façon qu'on ne peut priver du droit à la parole un non spécialiste du nucléaire, une personne qui n'est pas spécialiste de l'intelligence artificielle doit être considérée. Expliquons-lui et embarquons-la dans nos réflexions, car l'avenir de notre société est conditionnée par un partage des changements technologiques ».

Marc Watin-Augouard s'inscrit dans cette conception du partage du savoir : « Il n'y a pas de conscience sans formation, sans information, sans acculturation. Si l'on ne partage pas le savoir aujourd'hui, il y aura des ruptures. La fracture numérique n'est pas seulement technique et liée à l'accès au réseau, elle est aussi créée par l'incapacité qu'ont certaines personnes à comprendre leur environnement. La notion d'acceptabilité sociale me paraît très importante. Pour cette acceptabilité sociale, il faut que les citoyens soient apaisés par une approche pédagogique autre que celle peut-être qui domine aujourd'hui ».

C'est dans ce sens que s'exprime **Sawsen Rezig :** « On ne parle pas assez de formation, du travail pédagogique qu'il faut entreprendre auprès de la population, parce que les IA sont mises à disposition du grand public et que beaucoup de gens ne réalisent pas les conséquences, notamment dans le cas où, un jour peut-être, une IA aurait une "conscience". Il est donc indispensable de développer une action pédagogique vers le grand public pour améliorer l'acculturation de tous ».

...l'IA ressort aussi des sciences cognitives, des sciences humaines et sociales qui tentent d'expliquer ce qu'est l'humain dans son unicité, sa singularité.

L'IA **transformateur** de la société

L'IA, on l'a dit, a un impact très puissant sur l'humain, dans sa relation avec la machine, dans sa relation au savoir, dans sa manière de penser. Elle est aussi la cause de nombreuses transformations sociétales.

Je pense très clairement que les dirigeants d'entreprises doivent complètement repenser les modèles d'organisation et même les modèles de valeur.

Eric Salobir

Eric Salobir s'interroge sur une société optimisée par l'IA où l'on « serre les boulons » pour être toujours plus efficace, mais qui exploite toujours plus de ressources, etc. « Veut-on juste une société différente, au sein de laquelle nous allons travailler différemment ? Je pense très clairement que les dirigeants d'entreprises doivent complètement repenser les modèles d'organisation et même les modèles de valeur. En fait, la valeur ne va plus se trouver aux mêmes endroits et l'humain ne va pas apporter sa valeur au même endroit. Je pense qu'on sous-estime complètement la profondeur de la transformation des organisations qui est à venir ».

Guy-Philippe Goldstein identifie trois chocs pouvant impacter la société. Le premier concerne la cybersécurité ; il sera développé plus loin dans ce document. Le second est un choc bénéfique dans la manière d'organiser le travail dans la société : démocratisation voire « commoditisation[18] » de l'expertise, retour de l'échange socratique (encore plus clair avec le dialogue avec le *chatbot*), place renforcée à l'expérimentation continue de nouvelles capacités et, *in fine*, redéfinition de ce qui est essentiel pour le travail de l'utilisateur humain : la compréhension plus fine d'où se trouve la valeur. Avec l'IA s'opère une égalisation par le haut avec la transformation des organisations d'une structure pyramidale à une structure plate et égalitaire. **Guy-Philippe Goldstein** rejoint sur ce point la réflexion d'**Eric Salobir :** « C'est là où l'IA peut se révéler être une clé révolutionnaire en permettant à chacun d'avoir pour une somme modique son tuteur privé[19]. Le savoir implicite et explicite pourra directement vivre non pas dans la parole d'un sachant (avec le risque de hiérarchisation que cela implique) mais dans la réponse d'une ou plusieurs IA (dans un système dynamique, évolutif, critique avec plusieurs IA etc...).

18 Commoditisation, processus de transformation des produits ou services en objets standardisés et commercialisables. Ce processus a tendance à éliminer les qualités uniques ou distinctives de la marchandise au profit d'articles identiques à moindre coût qui peuvent être interchangeables entre eux.

19 Microsoft a nommé son IA *Copilot*, ce qui va dans le sens du tutorat.

D'un point de vue philosophique, c'est la poursuite de la révolution de l'écrit, démarrée il y a plus de 3 000 ans, où le texte remplace le prêtre comme dépositaire ultime du savoir sacré ». Le troisième choc s'opère en faveur des sociétés ouvertes ; c'est au niveau social la poursuite de l'opposition entre sociétés verticales et sociétés horizontales. L'IA aura deux variantes : celle qui favorise les structures horizontales (et potentiellement dangereuse pour les puissances autoritaires) mais qui donne l'augmentation la plus claire de productivité et l'IA des systèmes autoritaires qui sera tout aussi bonne d'un point de vue scientifique mais détournée, manipulée sur les questions de sciences sociales. On peut penser qu'à terme, comme pour la première guerre froide, c'est à nouveau le pôle capable de développer le plus d'échanges et d'esprits critiques/socratiques (grâce à une IA qui met à jour de plus en plus rapidement sa base de connaissance) qui remportera la guerre. Mais bien sûr, il y aura des tests car l'adversaire voudra instiller une culture du mensonge et/ou une culture de l'hallucination qui force à perdre pied et perdre confiance, un objectif clé du volet psychologique de la guerre hybride. La seconde guerre froide qui se jouera tout autant sur le terrain de l'unité nationale et la résilience sociale que celui de l'affrontement militaire direct.

Eric Salobir questionne l'IA au regard du projet de société : « Depuis des décennies, nous sommes habitués à subir de moins en moins de frictions. Le principe des technologies numériques est d'enlever cette friction. Nous vivons dans l'immédiat, dans la connexion automatique, etc. Et je pense que l'IA, c'est vraiment l'absolu de cela. Un certain nombre d'étudiants de Stanford, le soir, quand ils s'ennuient, discutent avec des *chatbots* ; le principe du *chatbot*, c'est qu'il est là pour vous dire exactement ce que vous avez envie d'entendre, vous emmener exactement là où vous souhaitez aller, vous «caresser dans le sens du poil». Au bout d'un moment, nous prenons conscience du principe d'une société, celui de se heurter à l'altérité, de se heurter à ceux qui ne sont pas d'accord. À un moment, nous risquons de nous déshabituer de ce principe et de trouver de plus en plus difficiles les interactions entre humains qui, par définition, sont un peu râpeuses, rugueuses. Cette rugosité fait que, petit à petit, l'humanité que l'on a en soi, on la reçoit aussi des autres. Notre humanité se forme à travers tout le personnalisme. C'est notre personnalité, elle se forme à travers le contact des autres. Donc pour moi la question est la suivante : quel projet de société veut-on bâtir ensemble ? Ensuite on adaptera la technologie. La technologie, c'est simplement un facilitateur (*enabler*) du projet de société. Je crains malheureusement qu'on n'ait pas vraiment réussi à se mettre d'accord sur un projet de société. Donc, finalement, la technologie ne nous entraîne que vers ce qui est parfois un peu notre penchant de facilité ».

Rayna Stamboliyska partage cette analyse : « Ce qui est réellement disrupté, c'est la société, ce sont nos interactions interpersonnelles, mais aussi notre rapport aussi à la vie en société. Parce que la libre expression, ce qu'on entend, c'est la garantie de l'expression d'un humain envers d'autres humains. Aujourd'hui, même sans aller jusqu'aux *fake news*, quand nous rencontrons une variabilité algorithmique en termes de production, de contenu, de campagne, là, ce ne sont plus des humains qui parlent à d'autres humains. Ce sont des humains qui donnent une idée vague, peut-être un peu caricaturée. Ce message est finalement élastique et adapté à la personne à qui il est destiné. La question qui se pose est la suivante : comment, nous, dans ce contexte-là, nous apprécions, nous qualifions et nous évoluons ? Qui porte finalement cette parole-là, celle qui traite de politique, de fabrique de la loi, de la création des règles de vie en société ? »

L'IA, on le comprend au travers des propos des intervenants, n'est pas une technologie – ou plus exactement un ensemble de technologies – qui produit un impact limité, sectoriel sur nos sociétés. Tous les paradigmes sont impactés. Dans ce contexte, la maîtrise de son développement est une nécessité. Cela exige de concevoir une « IA responsable ».

Ce qui est réellement disrupté, c'est la société, ce sont nos interactions interpersonnelles, mais aussi notre rapport aussi à la vie en société.

Rayna Stamboliyska

VERS UNE **IA RESPONSABLE ?**

Comme le soulignent très majoritairement les points de vue exprimés sur l'IA, cette technologie disruptive peut avoir des conséquences très importantes, voire très graves sur la vie de l'humanité. La crainte de voir émerger « l'apprenti sorcier », de voir dépassé un point de non-retour qui assujettirait l'humain, suscite une intense réflexion sur la finalité de l'IA, sur sa relation au bien et au mal. Une discipline qui relève dans sa conception, sa mise en œuvre, de sciences dites « dures » pénètre le champ du droit, de la sociologie, de la philosophie. Cette réflexion s'intensifie à mesure que les capacités de l'IA augmentent et bouleversent les paradigmes sur lesquels repose notre société. Elle est d'autant plus dominatrice dans le discours contemporain que chacun s'accorde à reconnaître les dégâts exceptionnels, inédits, que pourrait causer une IA non maîtrisée. « IA for good », IA pour le bien commun, tel est le dénominateur commun de toutes les démarches entreprises pour encadrer l'usage de l'IA. Les pionniers avaient déjà une approche éthique.

Une quête de **l'éthique**

La réflexion sur l'intelligence artificielle connaît une accélération depuis les années 2010. En septembre 2016, le Français Yann LeCun, un des principaux concepteurs du *deep learning*, mobilise les acteurs de la Big Tech[20] pour une initiative commune, *Partnership on Artificial Intelligence to Benefit People and Society.* L'objectif est de faire avancer la compréhension du public et définir les meilleurs pratiques sur les défis et les opportunités, parce que les participants sont convaincus que l'IA conduit à un changement de civilisation. « En ouvrant la conversation sur l'IA à une communauté plus large, nous espérons créer de nouveaux modèles d'engagement, de collaboration et de responsabilité pour faire avancer le domaine d'une manière réfléchie, positive et éthique qui profite aux personnes et à la société ».

Quelques mois plus tard, en janvier 2017, se tient la conférence d'Asilomar, en Californie. 2 000 signataires, dont l'astrophysicien Stephen Hawking et Elon Musk, patron de SpaceX, y adoptent un guide de référence comprenant 23 principes pour un développement éthique de l'intelligence artificielle.

20 Google, Amazon, Facebook, IBM, Microsoft puis Apple (2017).

LES 23 PRINCIPES D'ASILOMAR

1. **Objectif de ces recherches :** le développement de l'IA ne doit pas servir à créer une intelligence sans contrôle mais une intelligence bénéfique.

2. **Investissements :** les investissements dans l'IA doivent être soutenus par le financement de recherches visant à s'assurer de son usage bénéfique, qui prend en compte des questions épineuses en matière d'informatique, d'économie, de loi, d'éthique et de sciences sociales.

3. **Relations entre les scientifiques et les législateurs :** un échange constructif entre les développeurs d'IA et les législateurs est souhaitable.

4. **Esprit de la recherche :** un esprit de coopération, de confiance et de transparence devrait être entretenu entre les chercheurs et les scientifiques en charge de l'IA.

5. **Éviter une course :** les équipes qui travaillent sur les IA sont encouragées à coopérer pour éviter des raccourcis en matière de standards de sécurité.

6. **Sécurité :** les IA devraient être sécurisées tout au long de leur existence, une caractéristique vérifiable et applicable.

7. **Transparence en cas de problème :** dans le cas d'une blessure provoquée par une IA, il est nécessaire d'en trouver la cause.

8. **Transparence judiciaire :** toute implication d'un système autonome dans une décision judiciaire devrait être accompagnée d'une explication satisfaisante contrôlable par un humain.

9. **Responsabilité :** les concepteurs et les constructeurs d'IA avancées sont les premiers concernés par les conséquences morales de leurs utilisations, détournements et agissements. Ils doivent donc assumer la charge de les influencer.

10. **Concordance de valeurs :** les IA autonomes devraient être conçues de façon à ce que leurs objectifs et leur comportement s'avèrent conformes aux valeurs humaines.

11. **Valeurs humaines :** les IA doivent être conçues et fonctionner en accord avec les idéaux de la dignité, des droits et des libertés de l'homme, ainsi que de la diversité culturelle.

12. **Données personnelles :** chacun devrait avoir le droit d'accéder et de gérer les données le concernant au vu de la capacité des IA à analyser et utiliser ces données.

13. **Liberté et vie privée :** l'utilisation d'IA en matière de données personnelles ne doit pas rogner sur les libertés réelles ou perçues des citoyens.

14. **Bénéfice collectif :** les IA devraient bénéficier au plus de gens possible et les valoriser.

15. **Prospérité partagée :** la prospérité économique permise par les IA devrait être partagée au plus grand nombre, pour le bien de l'humanité.

16. **Contrôle humain :** les humains devraient pouvoir choisir comment et s'ils veulent reléguer des décisions de leur choix aux AI.

17. **Anti-renversement :** le pouvoir obtenu en contrôlant des IA très avancées devrait être soumis au respect et à l'amélioration des processus civiques dont dépend le bien-être de la société plutôt qu'à leur détournement.

18. **Course aux IA d'armement :** une course aux IA d'armement mortelles est à éviter.

19. **Avertissement sur les capacités :** en l'absence de consensus sur le sujet, il est recommandé d'éviter les hypothèses au sujet des capacités maximum des futures IA.

20. **Importance :** les IA avancées pourraient entraîner un changement drastique dans l'histoire de la vie sur terre, et doit donc être gérée avec un soin et des moyens considérables.

21. **Risques :** les risques causés par les IA, en particulier catastrophiques ou existentiels, sont sujets à des efforts de préparation et d'atténuation adaptés à leur impact supposé.

22. **Auto-développement infini :** les IA conçues pour s'auto-développer à l'infini ou s'auto-reproduire, au risque de devenir très nombreuses ou très avancées rapidement, doivent faire l'objet d'un contrôle de sécurité rigoureux.

23. **Bien commun : l**es intelligences surdéveloppées devraient seulement être développées pour contribuer à des idéaux éthiques partagés par le plus grand nombre et pour le bien de l'humanité plutôt que pour un État ou une entreprise.

La même année, le 3 novembre, l'université de Montréal lance les travaux préalables à la « Déclaration de Montréal pour une IA responsable », fruit du travail d'une équipe scientifique pluridisciplinaire et interuniversitaire. « Les principes de la présente déclaration reposent sur l'idée commune que les êtres humains cherchent à s'épanouir comme êtres sociaux doués de sensations, d'émotions et de pensées et qu'ils s'efforcent de réaliser leurs potentialités en exerçant librement leurs capacités affectives, morales et intellectuelles. Il incombe aux différents acteurs et décideurs publics et privés, au niveau local, national et international, de s'assurer que le développement et le déploiement de l'intelligence artificielle soient compatibles avec la protection et l'épanouissement des capacités humaines fondamentales ».

La déclaration fixe trois objectifs :

— Élaborer un cadre éthique pour le développement et le déploiement de l'IA

— Ouvrir un espace de dialogue national et international pour réussir collectivement un développement inclusif, équitable et écologiquement soutenable de l'IA

— Orienter la transition numérique afin que tous puissent bénéficier de cette révolution technologique

En décembre 2017, faisant suite à la loi pour une république numérique, la CNIL publie un rapport sur les enjeux éthiques de l'algorithme et de l'intelligence artificielle en identifiant six grandes problématiques :

— La délégation de tâches, de raisonnements et de décisions de plus en plus complexes et critiques à des machines

— Les risques de discrimination ou d'exclusion

— Les dangers du profilage et de la segmentation pouvant affecter l'individu comme les logiques collectives (pluralisme, mutualisation du risque)

— L'exploitation excessive de données à caractère personnel, au moment où le principe de minimisation est promu (*security by default*)

— Le nécessaire choix judicieux des données utilisées par les algorithmes d'apprentissage

— L'hybridation entre humains et machines, notamment dans les robots humanoïdes.

L'Union européenne s'inscrit dans ce mouvement. Telles sont les lignes de force des communications de la Commission de 2019 et 2020.

D'autres organisations internationales participent au mouvement. Ainsi les travaux de l'OCDE (mai 2019), de l'UNESCO (novembre 2021) soulignent la très grande convergence des acteurs. S'ouvre ainsi une dialectique entre éthique et innovation qui laisse entendre des cris d'inquiétude, lorsque certains prônent un moratoire dans le développement de l'IA. Ainsi, en 2021, Michelle Bachelet, la haut-commissaire des droits de l'homme de l'ONU évoque le risque grave d'atteinte aux droits de l'homme, car « les technologies d'intelligence artificielle peuvent avoir des effets négatifs, voire catastrophiques si elles sont utilisées sans prendre suffisamment en compte la manière dont elles affectent les droits humains ».

Récemment, plus de 1 300 chercheurs en intelligence artificielle (IA) appellent, le 29 mars 2023, à une « prise de recul » générale après le développement de ChatGPT-4. Dans cette « course dangereuse vers des boîtes noires imprévisibles », ils ont réclamé une pause de six mois dans la recherche, craignant des risques majeurs pour l'humanité : « Les systèmes AI dotés d'une intelligence humaine compétitive peuvent présenter des risques profonds pour la société et l'humanité, comme l'ont montré des recherches approfondies et ont été reconnus par les meilleurs laboratoires d'AI. Comme indiqué dans les principes d'Asilomar largement approuvés, *Advanced AI* pourrait représenter un changement profond dans l'histoire de la vie sur Terre ».

PRINCIPES DE L'UE

Communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions

Renforcer la confiance dans l'intelligence artificielle axée sur le facteur humain.

COM/2019/168 final reprise par la communication de la Commission européenne COM (2020) 65 du 19 février 2020 sur le *Livre Blanc Intelligence artificielle*, une approche européenne axée sur l'excellence et la confiance.

LIGNES DIRECTRICES ÉLABORÉES PAR LE GROUPE D'EXPERTS DE HAUT NIVEAU SUR L'IA

— **Facteur humain et contrôle humain :** les systèmes d'IA devraient être les vecteurs de sociétés équitables en se mettant au service de l'humain et des droits fondamentaux, sans restreindre ou dévoyer l'autonomie humaine

— **Robustesse et sécurité :** une IA digne de confiance nécessite des algorithmes suffisamment sûrs, fiables et robustes pour gérer les erreurs ou les incohérences dans toutes les phases du cycle de vie des systèmes d'IA

— **Respect de la vie privée et gouvernance des données :** il faut que les citoyens aient la maîtrise totale de leurs données personnelles et que les données les concernant ne soient pas utilisées contre eux à des fins préjudiciables ou discriminatoires

— **Transparence :** la traçabilité des systèmes d'IA doit être assurée

— **Diversité, non-discrimination et équité :** les systèmes d'IA devraient prendre en compte tout l'éventail des capacités, aptitudes et besoins humains, et leur accessibilité devrait être garantie

— **Bien-être sociétal et environnemental :** les systèmes d'IA devraient être utilisés pour soutenir des évolutions sociales positives et renforcer la durabilité et la responsabilité écologique

Ces appels trouvent un écho dans l'organisation du forum inaugural AI Insight, organisé par le Sénat américain, au Capitole, le 13 septembre 2023, qui a rassemblé de nombreux acteurs de la Big Tech, dont Sundar Pichai, Elon Musk, Mark Zuckerberg, Bill Gates. Dans cette mouvance, intervient l'ordonnance de Joe Biden du 30 octobre 2023 pour une IA sûre, sécurisée et digne de confiance.

Quelques jours après, le gouvernement britannique de Rishi Sunak organise, les 1er et 2 novembre 2023, à Bletchley Park, la première conférence mondiale consacrée aux risques associés à l'intelligence artificielle (IA).

Le Premier ministre britannique veut proposer la création d'un groupe d'experts internationaux, sur le modèle du GIEC, chargé de publier un état des lieux de l'IA.

L'ONU, de son côté, s'exprime, notamment par la voix de son secrétaire général, Antonio Gutteres, lors du Conseil de sécurité de l'ONU, le 18 juillet 2023.

Celui-ci a averti que si l'intelligence artificielle devenait principalement une arme pour lancer des cyberattaques, générer des *deepfakes*, ou pour diffuser de la désinformation et des discours de haine, cela aurait des conséquences très graves pour la paix et la sécurité mondiales.

« L'IA doit profiter à tous, y compris au tiers de l'humanité qui se trouve encore hors ligne », déclare-t-il, lors du sommet mondial *AI for Good* (IA pour le bien), organisé les 6 et 7 juillet 2023, par l'Union internationale des télécommunications (UIT) à Genève.

Le 26 octobre 2023, le Secrétaire général de l'ONU annonce la création d'un nouveau Conseil consultatif sur l'intelligence artificielle sur les risques, les opportunités et la gouvernance internationale de l'intelligence artificielle. Cet organisme doit soutenir les efforts de la communauté internationale pour encadrer l'intelligence artificielle. Plus récemment, l'Assemblée générale des Nations Unies adopte, le 21 mars 2024, une résolution intitulée « Saisir les possibilités offertes par des systèmes d'intelligence artificielle sûrs, sécurisés et dignes de confiance pour le développement durable ».

« Les systèmes d'intelligence artificielle – lit-on – qui ne relèvent pas du domaine militaire, suivent un cycle de vie passant par les étapes de la préconception, de la conception, de la mise au point, de l'évaluation, de la mise à l'essai, de la mise en service, de l'utilisation, de la vente, de l'achat, de l'exploitation et de la mise hors service. Ils sont centrés sur l'être humain, fiables, explicables, éthiques et inclusifs et pleinement ancrés dans le respect, la promotion et la protection des droits humains et du droit international ; ils veillent au respect de la vie privée, sont axés sur le développement durable et sont responsables. Ils ont le potentiel de favoriser la transformation numérique, de promouvoir la paix, de combler le fossé numérique entre les pays et à l'intérieur même des pays et de favoriser et de garantir la jouissance des droits humains et des libertés fondamentales par tous, en veillant à ce que la personne humaine conserve sa place centrale ». **Benoît Tabaka**, lors de l'Agora, souligne que les 17 objectifs fixés par l'ONU seront loin d'être atteints en 2030.

Enfin, lors de la 133e session du Comité des ministres, le Conseil de l'Europe adopte, le 17 mai 2024, la convention-cadre du Conseil de l'Europe sur l'intelligence artificielle, premier traité international juridiquement contraignant visant à garantir le respect des normes juridiques en matière de droits de l'homme, de démocratie et d'État de droit dans le cadre du recours aux systèmes d'intelligence artificielle (IA).

Cet aperçu historique incomplet des prises de conscience et des initiatives nous font mieux comprendre le titre du rapport de Cédric Villani (2018)[21], « Donner un sens à l'intelligence artificielle ». Il s'agit bien de donner du sens, c'est-à-dire de mettre la technologie en perspective en développant une éthique collective : « Si nous souhaitons faire émerger des technologies d'IA conformes à nos valeurs et normes sociales, il faut agir dès à présent en mobilisant la communauté scientifique, les pouvoirs publics, les industriels, les entrepreneurs et les organisations de la société civile. Notre mission a cherché, modestement, à proposer quelques pistes permettant de poser les bases d'un cadre éthique pour le développement de l'IA et à faire vivre ce débat dans la société ».

> L'IA doit profiter à tous, y compris au tiers de l'humanité qui se trouve encore hors ligne.

21 Cédric VILLANI, « Donner un sens à intelligence artificielle », rapport précité, 2018.

Le mythe de **l'IA responsable**

L'IA peut-elle être responsable ? Qui est responsable de l'IA ? La question de la responsabilité appelle une approche juridique qui ne peut totalement s'appuyer sur le corpus existant. Les nouvelles technologies précèdent toujours le droit qui vient postérieurement encadrer leurs effets.

Le droit distingue l'humain, les animaux (donc le vivant), des objets matériels et immatériels (un logiciel, une œuvre de l'esprit). L'humain est responsable de son propre fait et des choses qu'il a conçues, qu'il détient, dont il a la garde ou qu'il met en œuvre. Ainsi, l'article 1240 du Code civil des Français dispose que « Tout fait quelconque de l'homme, qui cause à autrui un dommage, oblige celui par la faute duquel il est arrivé à le réparer ». Le régime juridique de la responsabilité a connu une extension avec l'implication des personnes morales, dont la responsabilité civile ou pénale peut être mise en cause.

Les robots peuvent être immatériels, comme ils peuvent avoir une apparence physique. L'intelligence artificielle qui les anime peut avoir un degré d'autonomie avancé, sans supervision humaine. Dès lors, faut-il leur reconnaître une personnalité juridique, emportant des droits et des obligations, qui s'ajouterait à celle des personnes physiques et des personnes morales ?

Le Parlement européen, par sa résolution du 16 février 2017, pose le problème en ces termes : « Plus un robot est autonome, moins il peut être considéré comme un simple outil contrôlé par d'autres acteurs (tels que le fabricant, l'opérateur, le propriétaire, l'utilisateur, etc.) ; à cet égard se pose la question de savoir si les règles ordinaires en matière de responsabilité sont suffisantes ou si des principes et règles nouveaux s'imposent pour clarifier la responsabilité juridique des divers acteurs en matière de responsabilité pour les actes ou l'inaction d'un robot dont la cause ne peut être attribuée à un acteur humain en particulier, et pour déterminer si les actes ou l'inaction du robot qui ont causé des dommages auraient pu être évités ». La résolution ajoute que « dans l'hypothèse où un robot puisse prendre des décisions de manière autonome, les règles habituelles ne suffiraient pas à établir la responsabilité juridique pour dommages causés par un robot, puisqu'elles ne permettraient pas de déterminer quelle est la partie responsable pour le versement des dommages et intérêts ni d'exiger de cette partie qu'elle répare les dégâts causés ». Ceci est d'autant plus nécessaire que la résolution admet « qu'il est possible, en fin de compte, qu'à long terme, l'intelligence artificielle surpasse les capacités intellectuelles de l'être humain ». La question de la responsabilité des robots porte sur l'attribution du dommage causé à une personne ou à un bien. Dans le droit actuel, la responsabilité incombe généralement au fabricant du robot ou à son propriétaire. Toutefois, cette approche peut s'avérer insuffisante à mesure que les robots acquièrent une autonomie et une capacité de décision de plus en plus grandes. Mais le Parlement européen revient à la notion de responsabilité des personnes physiques, car « la tendance à l'automatisation demande que les personnes participant au développement et à la commercialisation des

applications de l'intelligence artificielle y intègrent la sécurité et l'éthique dès le départ, et reconnaissent ainsi qu'elles doivent être prêtes à accepter la responsabilité juridique de la qualité de la technologie qu'elles produisent ». On notera que le règlement européen sur les machines (2023/1230) du 29 juin 2023 remplace la directive machines (2006/42/CE). Il fait suite au rapport du 19 février 2020 de la Commission sur les conséquences de l'intelligence artificielle, de l'Internet des objets et de la robotique sur la sécurité et la responsabilité. Il intègre les nouveaux risques engendrés par l'intelligence artificielle, sans remettre en cause le régime juridique de la responsabilité.

L'idée d'une responsabilité civile ou pénale des robots exigerait la révision de notions fondamentales du droit civil ou pénal qui conduirait à juger la technologie et non la personne physique qui la met en œuvre. Reconnaître que l'IA peut avoir une « conscience », une capacité de discernement, créerait un véritable séisme juridique.

Reconnaître que l'IA peut avoir une « conscience », une capacité de discernement, créerait un véritable séisme juridique.

L'humain est **responsable** de l'IA

Jean-Philippe Desbiolles et **Grégoire Colombet** excluent de tenir l'IA pour responsable : « L'IA ne dispose pas à ce jour de conscience, de capacité à éprouver des choses subjectivement qui la rendrait assujettie aux mêmes lois que les personnes physiques ou morales. Ils opèrent une distinction selon que la décision préjudiciable a été prise en autonomie (dans ce cas le concepteur de l'IA est irresponsable) ou qu'elle a été influencée. Concernant cette dernière hypothèse, il faudrait incomber la responsabilité des mauvaises décisions aux détenteurs des systèmes utilisant le modèle d'IA qui n'a pas su maintenir l'autonomie de ses utilisateurs face à l'intelligence artificielle *».*

Grégoire Colombet ajoute : « On peut prendre autant de décisions que l'on veut à un moment donné avec un système algorithmique ; à la fin, il y aura toujours un humain qui est responsable. On parle de conscience. Il faut avoir les moyens de cette conscience parce que, finalement, ce dont on s'est aperçu, c'est que la collaboration entre l'humain et la machine n'est pas évidente. Il faut préserver notre libre arbitre dans la manière dont on fait collaborer des utilisateurs avec ces systèmes algorithmiques, parce qu'on est influencé, parce qu'on se laisse entraîner par la machine ».

Catherine Morin Dessaily s'inscrit dans cette approche : « Ce serait dangereux de considérer que l'IA est responsable. Ça voudrait dire qu'elle serait volontaire et autonome de ses décisions. Nous sommes responsables de l'IA et devons fixer les limites aux mésusages. »

Avec le regard de l'universitaire, **Michel Sejean** considère comme un abus de langage qui ne trompe pas l'expression « IA responsable ». Une expression malheureuse de même nature a déjà été employée à propos du développement durable. « Ce n'est pas le développement lui-même qui est durable ; il doit être conciliable avec son environnement, de manière que son environnement soit durable. L'IA «responsable» doit être conciliable avec son environnement, de manière que son environnement reste responsable. Il y a, c'est vrai, des plaideurs qui ont essayé de défendre le contraire. Une affaire concernant une compagnie aérienne illustre parfaitement cette tentation de dérive : un voyageur avait dû annuler son billet qui, dans la politique d'annulation de la compagnie, n'était pas remboursable. Mais il est allé consulter le *chatbot* qui est en bas à droite de la page du site officiel qui dit : « Bonjour, je m'appelle Nina, je suis votre *chatbot* ». Le client lui expose sa situation : « Pensez-vous que c'est remboursable? ». Le *chatbot* lui répond par l'affirmative. Fort de cette réponse, il se retourne vers la compagnie qui ne confirme pas les propos de son *chatbot* et refuse de rembourser parce que, dit-elle, « notre IA est autonome. Donc nous n'en sommes pas responsables ». Les tribunaux ont tranché : « Non, l'IA n'est pas autonome au sens juridique. Elle n'est pas sa propre règle autonome. C'est vous qui avez demandé à cette IA de répondre. Vous en êtes responsable ».

Rayna Stamboliyska juge que le mot responsabilité n'est pas suffisant : « On parle surtout de *accountability*, en anglais, ou plutôt de redevabilité. C'est très bien d'être responsable, mais à un moment donné, si on est coupable, que se passe-t-il ? Si nous n'avons pas de mécanisme de compte-rendu, donc de transparence, on peut disserter sur la responsabilité, mais nous n'irons pas plus loin. Donc la question, quand on parle de la responsabilité à propos des algorithmes, est la suivante : peut-on le faire, doit-on le faire ? La question porte surtout sur le « comment ? ». Aujourd'hui, je ne suis pas du tout convaincue qu'on ait dépassé le stade de la question ».

Cette réflexion élargit le champ du débat : il faut réguler, ce qui implique bien sûr de réglementer. **Benoît Tabaka** n'est pas d'un avis différent lorsqu'il déclare que l'IA est une technologie aujourd'hui trop importante pour ne pas être régulée. « L'intérêt, c'est d'intégrer l'intégralité de la chaîne de valeur, depuis celui qui développe les grands modèles de langage, en quelque sorte les éléments fondateurs utilisés par différents acteurs jusqu'à la fin, le développeur de l'application qui intègre des modèles de langage pour des usages très particuliers. Aujourd'hui, votre modèle de langage, que ce soit celui de Google comme Gemini, que ce soit Mistral, ou celui d'Aleph Alpha, etc., ce modèle va se retrouver dans des multiples applications ; des applications qui peuvent être très importantes, nécessaires à la science, à la santé, à l'environnement, mais aussi qui peuvent être néfastes, avec des outils pour développer des *deepfakes* ou pour lancer des cyberattaques. Et c'est là où la responsabilité de l'ensemble de la chaîne est importante. Il faut faire peser des obligations sur ceux qui vont mettre en place ces fameux larges modèles de langage et notamment avec de la transparence, avec de "l'accountabilité", avec de la documentation. Un développeur qui met en service un outil de génération d'images doit intégrer des outils de *watermarking*[22], pour s'assurer que les contenus qui seront produits pourront être tracés ensuite sur les réseaux ».

Les intervenants sont unanimes : l'IA n'est pas responsable ; les humains sont responsables des conséquences de la mise en œuvre de l'IA. Pour l'heure, il n'est pas question d'envisager de doter l'IA d'une personnalité juridique, hypothèse très contestée de la résolution du Parlement européen de 2017. Il y a lieu de s'en tenir à la conception classique qui place l'humain au cœur de la notion de responsabilité. La résolution du Parlement européen du 16 février 2017, contestée par certains aspects, est claire sur ce point : « La tendance à l'automatisation demande que les personnes participant au développement et à la commercialisation des applications de l'intelligence artificielle y intègrent la sécurité et l'éthique dès le départ, et reconnaissent ainsi qu'elles doivent être prêtes à accepter la responsabilité juridique de la qualité de la technologie qu'elles produisent ». Telle est la position que l'on doit adopter aujourd'hui, sans rejeter toute réflexion prospective, eu égard aux évolutions potentielles de l'IA, notamment si elle atteint le stade de l'intelligence artificielle générale, censée égaler, voire dépasser l'intelligence humaine, notamment sur certains aspects comme la célérité de calcul, de traitement, d'exécution, ou par la capacité d'anticipation.

C'est dans cet esprit que s'inscrit l'*AI Act*, fruit d'un accord obtenu après de longs échanges entre la Commission, le Conseil de l'Union européenne et le Parlement.

Il faut faire peser des obligations sur ceux qui vont mettre en place ces fameux larges modèles de langage et notamment avec de la transparence, avec de «l'accountabilité», avec de la documentation.

Benoît Tabaka

22 *Watermarking* : « tatouage numérique », technique qui consiste à insérer un logo, un texte ou un symbole sur une image pour en indiquer la source ou le propriétaire.

L'*AI Act*

Miguel Ángel Cañada rappelle que les législations actuelles en matière d'IA jusqu'à l'*IA Act* ne sont pas adaptées. « Aujourd'hui, il faut prendre en compte deux choses. D'abord, la gestion du risque et la deuxième chose n'est pas une question technique, mais une question philosophique qui concerne la manière dont on comprend cette intelligence artificielle. En qualité d'Européens, nous devons être les champions de cela dans le monde. La technologie va être différente de celle que nous avons gérée au cours des dernières années. Il nous faut gérer le risque. Toutes les législations, toutes les approches doivent aller dans cette direction et cela fait partie de la réflexion que nous faisons à la Commission européenne. Nous allons devoir vivre avec et nous devons vivre avec ça et changer notre état d'esprit dans les années à venir ».

La réglementation sur l'IA s'inscrit dans une tendance à l'universalité, comme le souligne Michel Sejean : « En droit, à l'échelle de la planète, les solutions juridiques sont souvent très différentes. Je ne sais pas s'il y a un équivalent dans l'histoire du droit, mais c'est un cas où les instruments internationaux sont en train de récupérer la même définition de l'IA, celle de l'OCDE, qui est vraiment en train de faire autorité, d'avoir une portée qui se veut universelle. Je pense au Conseil de l'Europe, organisation internationale à 46 membres, qui vient de publier un projet de convention cadre sur l'intelligence artificielle dans sa compatibilité avec les droits humains. Exemple, si je suis en procès, je dois pouvoir décider que ce n'est pas une machine qui va juger de mon sort, je dois donc être au courant que c'est une machine qui est dans le processus, et je dois pouvoir la mettre de côté et dire que c'est un humain qui va régler mon sort. Cette convention, au lieu de s'appliquer aux 46 États membres, a vocation à s'appliquer à tous les États, même ceux qui ne sont pas membres

La Convention sur l'IA du Conseil de l'Europe a vocation à s'appliquer à tous les États. C'est sans doute une première dans l'histoire du droit.

Michel Sejean

du Conseil de l'Europe, qui voudraient s'inspirer de cette convention. L'inquiétude se manifeste dans tous les pays du monde, de la même manière, alors que d'habitude, on a des approches différentes selon les systèmes juridiques. C'est sans doute une première dans l'histoire du droit et c'est vraiment remarquable ».

Sandrine Hilaire est moins affirmative : « Une IA responsable chinoise, une IA responsable américaine et une IA responsable française. Je vous laisse voir, ce n'est pas la même culture, pas la même manière d'appréhender les choses, pas les mêmes logiques juridiques, pas du tout les mêmes volontés sociétales. Est-ce qu'on va arriver à quelque chose de général dans tout ça, de responsable pour tout le monde ? »

Pour **Mireille Clapot**, l'idéal serait une règlementation universelle dans le cadre de l'ONU, à l'instar de ce qui se pratique pour l'aviation (OACI) ou le domaine maritime (Convention de Montego Bay). Mais déjà à l'échelon européen, « l'*AI Act* permet d'avoir, en fonction des valeurs européennes, un certain nombre de régulations. C'est le bon équilibre entre la protection des valeurs, de l'État de droit, de la liberté d'expression, liberté fondamentale, et l'innovation. Les deux grandes puissances ont choisi l'une de tout innover au détriment des libertés et l'autre de tout réguler au détriment aussi des libertés ».

Grégoire Colombet souhaite que la quête de performance s'inspire des idéaux partagés dans une vision éthique de l'AI. L'*AI Act* répond à cette attente, mais sa philosophie sera-t-elle partagée à l'échelle de la planète ? Sa question est pertinente, au regard du contexte géopolitique, mais il est probable que l'*AI Act* aura des effets universels, comme ceux que l'on observe avec le RGPD, sous l'influence des décisions d'adéquation.

Emmanuelle Legrand, ancienne négociatrice française du règlement, rappelle que « l'approche qui a été préconisée par la Commission n'est pas de réguler une technologie qui n'est d'ailleurs pas unique, puisqu'il y a plusieurs technologies. L'objectif de l'*AI Act* est de réguler l'usage de ces technologies, notamment celles à hauts risques et, dans certains cas, d'en interdire certains usages déclarés incompatibles avec les valeurs de l'Union européenne. Puisqu'il y a un humain derrière la machine, l'*AI Act* consiste aussi à réguler les comportements humains. Le règlement IA ne concerne pas que l'IA générative qui n'était pas visée à l'origine mais a été intégrée, notamment au moment de la présidence française de l'UE et bien avant la sortie commerciale de ChatGPT, dans le cadre des systèmes dits à usage générique. L'*AI Act* est une législation qui aborde l'IA non comme une technologie mais comme un système qui est mis sur un marché, ce qu'on appelle une législation-produit. Il n'avait pas vocation – et cela a été assumé jusqu'au trilogue avec le Parlement – à inclure et à régler tous les problèmes, notamment les problèmes environnementaux. Il n'avait pas vocation non plus à inclure ce qui était lié au droit d'auteur, simplement du fait de l'approche qui avait été choisie par la Commission européenne et que les États membres avaient maintenu dans le cadre du Conseil ».

Catherine Morin-Desailly apprécie l'approche de l'Union européenne qui consiste à réguler par le risque. Il y a des intelligences artificielles, en tout cas des applications qui ne peuvent pas être autorisées parce qu'elles vont porter très gravement atteinte à tous nos droits fondamentaux. Celles qui, par exemple, vont viser à modifier nos comportements, les influencer de manière subliminale. L'ONU a d'ailleurs, comme le rappelle la sénatrice, voté une résolution en ce sens, le 21 mars 2024.

Puisqu'il y a un humain derrière la machine, l'*AI Act* consiste aussi à réguler les comportements humains.

Emmanuelle Legrand

« Considérant que les systèmes d'intelligence artificielle sûrs, sécurisés et dignes de confiance – à savoir, pour les besoins de la présente résolution –, les systèmes d'intelligence artificielle qui ne relèvent pas du domaine militaire, suivent un cycle de vie passant par les étapes de la préconception, de la conception, de la mise au point, de l'évaluation, de la mise à l'essai, de la mise en service, de l'utilisation, de la vente, de l'achat, de l'exploitation et de la mise hors service, sont centrés sur l'être humain, fiables, explicables, éthiques et inclusifs et pleinement ancrés dans le respect, la promotion et la protection des droits humains et du droit international, veillent au respect de la vie privée, sont axés sur le développement durable et sont responsables ».

Le législateur européen a choisi une approche « horizontale ». Il aurait pu réglementer de manière « verticale », secteur par secteur. Il a préféré un dominateur commun : le risque. L'*AI Act* aborde la responsabilité avec une approche par le risque en distinguant les systèmes d'intelligence artificielle (SIA) à risque inacceptable, les SIA à haut risque, les SIA à risque faible et les SIA à risque minime.

Les IA créant des risques inacceptables sont interdites (systèmes de notation sociale, l'IA manipulatrice, altération substantielle des comportements humains pouvant créer un préjudice psychologique ou physique).

La majeure partie du règlement porte sur les systèmes d'IA à haut risque, qui sont réglementés. Les systèmes d'IA à risque limité sont soumis à des obligations plus légères (transparence permettant aux utilisateurs de savoir qu'ils interagissent avec une IA).

Le règlement *AI Act* s'applique aux personnes, fournisseurs, « déployeurs » (toute personne physique ou morale, autorité publique, agence ou autre organisme utilisant un système d'IA sous son autorité), importateurs, distributeurs et fabricants de systèmes d'intelligence artificielle (SIA), personnes physiques ou morales, dont le siège est situé sur le territoire européen ou dont les produits sont commercialisés sur le marché européen.

L'*AI Act* comme le règlement sur la cyber-résilience (CRA) responsabilise les acteurs du numérique en veillant à ce que « les fabricants prennent la sécurité au sérieux tout au long du cycle de vie d'un produit ». Le texte est particulièrement exigeant en termes de responsabilité s'agissant de l'opérateur d'un système d'IA, notamment à haut risque, qui est objectivement responsable de tout préjudice ou de tout dommage causé par une activité, un dispositif ou un procédé physique ou virtuel piloté par un système d'IA.

Le projet de Directive européenne relative à la responsabilité en matière d'intelligence artificielle (AILD)[23] modifie le droit de l'UE en matière de responsabilité civile, en introduisant pour la première fois des règles spécifiques aux dommages causés par des systèmes d'IA. La directive introduit deux mesures principales : la « présomption de causalité », qui dispensera les victimes de l'obligation d'expliquer en détail comment le dommage a été causé par une faute ou une omission spécifique et l'accès aux éléments de preuve détenus par les entreprises ou les fournisseurs, lorsque ces derniers utilisent de l'IA à haut risque.

23 Proposition de Directive du Parlement européen et du Conseil relative à l'adaptation des règles en matière de responsabilité civile extracontractuelle au domaine de l'intelligence artificielle (Directive sur la responsabilité en matière d'IA). COM/2022/496 final.

FOCUS SUR L'*AI ACT*

IA ET CYBERSÉCURITÉ

Il n’est pas possible d’aborder le sujet de l’IA sans un regard sur la cybersécurité, condition sine qua non de la confiance. Les développements qui précèdent se rapportent, parfois indirectement, à une conception lato sensu de la cybersécurité qui protège les internautes, notamment au travers des contenus qui peuvent être illicites, des atteintes à la confiance. Stricto sensu, la cybersécurité porte sur l’intégrité des systèmes, des données et sur la lutte contre la cybercriminalité. Quelles sont, dans une approche large ou plus resserrée, les interactions entre l’IA et la cybersécurité ?

Nous sommes dans une dialectique des voleurs et des policiers, où les voleurs, potentiellement, ont peut-être un « risque d’avance.

Guy-Philippe Goldstein

Guy-Phillipe Goldstein rappelle que l’IA est le premier risque identifié chez les avocats américains, chez Gartner. « La « démocratisation » de certaines techniques d’attaques à un niveau plus ou moins sophistiqué (*phishing*, scan des vulnérabilités, *reverse engineering*) pourrait permettre à un spectre plus large d’individus de tenter une carrière cybercriminelle. Nous observons déjà une augmentation de 1200 % des attaques de *phishing* sur un an. En outre, certaines des propriétés émergentes des LLMs peuvent permettre de manipuler ces systèmes par le nouveau langage de programmation désormais le plus à la mode : l’Anglais/le Français/le langage naturel. Comme d’habitude, nous sommes dans une dialectique des voleurs et des policiers, où les voleurs, potentiellement, ont peut-être un « risque d’avance ». Ce risque pourrait être contrebalancé en retour par une démocratisation plus large du nombre de défenseurs, en même temps qu’une automatisation forte de plus en plus de tâches, libérant du temps pour pouvoir se former. Nous sommes donc dans une course de vitesse. Si nous restons sur les modèles du passé, les assaillants auront une longueur d’avance. Rien n’est cependant perdu pour les défenseurs, mais cela exige qu’ils soient capables de réinvestir une partie des plus-values obtenues par l’automatisation supplémentaire dans la formation et la réorganisation de l’action des collaborateurs. Les *business owners* pourront et devront prendre en compte la sécurité dès la conception (*security by design*). Sinon il y aura un choc dans la société car l’IA va devenir une couche essentielle dans l’ensembles des technologies permettant de faire fonctionner un *process* dans l’entreprise, la « *stack* informatique », elle-même déjà une couche essentielle de la création de richesse, et cela va s’accélérer.

La cybersécurité représente déjà un coût très important. À l’ère de l’IA-cybersécurité, ce coût, s’il n’est pas maitrisé, pourrait être exorbitant ».

Son métier, c’est la cartographie de la cybercriminalité, donc une analyse massive de données. **Jérôme Clauzade** souligne l’intérêt de l’IA par sa capacité à classifier des informations. « Elle fait gagner un temps incroyable dans l’analyse de corrélations que l’on n’aurait pas imaginé pouvoir faire manuellement. Nous avons aussi beaucoup progressé sur l’entraînement de nos équipes, c’est-à-dire acquérir du savoir, trouver facilement une information. C’est encore mieux qu’un moteur de recherche. C’est beaucoup plus performant. Le problème, c’est qu’on n’est pas les seuls à avoir cette idée. Il y a des gens qui sont moins bien intentionnés que nous qui ont fait la même chose que nous. Parce que les IA sont accessibles, ils peuvent détourner les IA et LLM à des fins mal intentionnées et les plus ambitieux d’entre eux peuvent créer leur propre LLM. Au final, ce qu’on obtient, ce sont des gens qui ne sont pas plus brillants qu’avant mais qui sont juste beaucoup plus efficaces qu’avant dans le volume et la pertinence de leurs attaques. Ce que nous avons introduit dans la cybercriminalité avec l’IA, c’est une notion de productivité. Certains ont entraîné leurs équipes à être beaucoup plus performantes, à être beaucoup plus pertinentes, plus rapidement, et surtout à moindre coût. On leur a appris à être productifs. Un exploit, une vulnérabilité coûte très cher à trouver. Il faut donc rentabiliser l’investissement quand les groupes en acquièrent. Grâce à l’IA, les prédateurs ne seront pas plus performants qu’avant en termes de qualité d’attaque ; ils vont être plus efficaces et surtout opérer à une échelle qui est sans commune mesure ».

Grâce à l'IA, les prédateurs ne seront pas plus performants qu'avant en termes de qualité d'attaque ; ils vont être plus efficaces et surtout opérer à une échelle qui est sans commune mesure.

Jérôme Clauzade

« Il y a une dernière étape, peut-être inquiétante pour le reste du monde, s'inquiète **Jérôme Clauzade**, lorsque les IA vont être capables de concevoir des choses auxquelles on n'avait pas encore pensé. Ce n'est pas le cas en ce moment, car les IA sont des robots. Ce sont des robots stupides qui reproduisent juste plus vite ce qu'on sait faire, en tout cas en termes de cybersécurité. Nous n'apprenons rien d'une IA. Nous n'avons rien découvert grâce aux IA pour le moment. Nous allons juste plus vite à moindre coût. Plus précisément, le coût est moindre pour les prédateurs.

Plus inquiétant est ce qui va arriver avec la prochaine génération d'IA qui va permettre aux prédateurs de concevoir de nouvelles techniques d'attaque en utilisant la capacité d'attaque « brute force » massive des IA et les combiner avec leur propre capacité de *social engineering*, pour créer des faux profits LinkedIn, se faire passer pour des faux collaborateurs, obtenir des informations qui vont rendre les attaques encore plus pertinentes. Donc ça, c'est demain. Nous nous y préparons. Nous sommes un peu inquiets ».

Il est essentiel de reconnaître que certains pays, qui n'adoptent pas la même approche que nous, développent intensivement l'intelligence artificielle sans suivre une trajectoire éthique.

Miguel Ángel Cañada

Cette préoccupation est également partagée par **Miguel Ángel Cañada**, qui redoute une réglementation asymétrique : « Il est essentiel de reconnaître que certains pays, qui n'adoptent pas la même approche que nous, développent intensivement l'intelligence artificielle sans suivre une trajectoire éthique. Ils adoptent une stratégie de déstabilisation et de confrontation. Ainsi, il est crucial de prendre en compte tous ces développements extraordinaires en matière d'intelligence artificielle, actuels et futurs. Les meilleures recherches et les développements les plus avancés se trouvent parfois dans le mauvais camp. La cybercriminalité est devenue un secteur très lucratif, générant des revenus bien supérieurs à ceux des organisations criminelles traditionnelles. Les cybercriminels cherchent constamment à accroître leurs gains. Par conséquent, il est impératif de contrer ces comportements malveillants en développant des technologies qui nous permettent de comprendre et d'anticiper les menaces potentielles. Ces acteurs malintentionnés continueront d'investir massivement sans se soucier de notre législation. Je rejoins ceux qui insistent sur la nécessité de protéger nos sociétés. Si nous ne le faisons pas, d'autres le feront, et nous ne serons pas prêts à affronter les défis posés par leur maîtrise de ces technologies. Il est donc essentiel d'être extrêmement vigilants et attentifs. Le besoin de cybersécurité est une réalité pressante. Nous devons intensifier nos investissements dans ces technologies car nos adversaires le feront sans considération pour les lois ou les responsabilités. De plus, certains pays pourraient utiliser ces technologies pour déstabiliser d'autres nations. Assurément, il est crucial de réglementer l'utilisation de l'IA et d'insister sur l'importance de l'éthique dans son déploiement.

Il est également vital d'accroître la coopération internationale, non seulement pour l'échange d'informations mais aussi pour la recherche, car nous ne disposons pas des mêmes ressources que nos adversaires ».

Jérôme Clauzade partage cette analyse : « Il faut prendre en compte aussi le fait que ceux qui nourrissent des IA offensives n'ont pas de règles, ne respectent aucun code, ont accès à des données globales des challenges *Capture the Flag* (CTF), des vulnérabilités, des bases de connaissances. Les prédateurs sont très volubiles et partagent beaucoup d'informations, des données qu'ils ont récupérées de manière frauduleuse. Ils peuvent utiliser massivement ces données, sans foi ni loi. En revanche, de notre côté, nous ne pouvons pas utiliser les données métiers des gens que nous souhaitons protéger. Si je demande à l'ensemble des banques du monde de partager avec moi leurs données, surtout leurs données métiers, pour m'aider à entraîner une IA pour les protéger, je vais me heurter d'abord à un refus, forcément, et surtout parce qu'on est tous protégés par le RGPD, par cet ensemble de lois qui nous gouverne, qui protège nos données de citoyens.

Donc, nous n'avons pas accès au même « terrain de jeu », car nous n'avons pas accès à toutes ces informations. Nous devons innover, c'est-à-dire apprendre localement plutôt qu'apprendre globalement. Nous n'avons pas la même liberté. Donc, pour garder l'avantage, nous devons être un peu plus malins ».

Cette asymétrie décrite ci-dessus est bien observée dans la pratique quotidienne, comme le souligne **Marc Watin-Augouard** : « L'*AI Act* encadre très strictement le recours à l'IA par les forces de sécurité intérieure mais est muet s'agissant des prédateurs. L'encadrement des modes d'action des enquêteurs peut se comprendre dans une démocratie, à condition qu'il n'y ait pas de déséquilibre. Au-delà des policiers et des gendarmes se pose la question de la victime qui n'est jamais citée dans la jurisprudence de la Cour de justice de l'Union européenne, toujours prompte à restreindre les conditions de recours aux investigations numériques pour protéger les libertés, notamment la vie privée. Or, la première des libertés, c'est la sécurité, c'est la protection des victimes. Faute de capacités d'action, le nombre de ces victimes pourrait singulièrement augmenter. La recherche d'un équilibre sécurité-liberté est un des enjeux qui domine la cybersécurité, aujourd'hui, mais surtout demain ».

Il est important de continuer l'effort de formation dans le domaine de l'IA pour comprendre ces enjeux et ces liens qui vont être créés entre la cybersécurité et l'IA.

Thiébaut Meyer

Selon **Thiébaut Meyer,** la technologie sera accessible aux attaquants comme aux défenseurs. Mais, plus optimiste que **Miguel Ángel Cañada** et **Jérôme Clauzade,** il pense qu'il y a un petit avantage côté défenseurs. « La simple raison, c'est que nous pourrons entraîner ces applications sur des données et des contextes qui sont propres ; on a parlé de *fine-tuning* des modèles, c'est-à-dire de leur réglage fin. Ensuite, il est important de continuer l'effort de formation dans le domaine de l'IA pour comprendre ces enjeux et ces liens qui vont être créés entre la cybersécurité et l'IA. Enfin, en termes de réglementation, de régulation, on voit que les grands progrès en termes de sécurité ont été possibles sur la notion de sécurité ouverte, sur des standards partagés pour qu'on puisse augmenter le niveau de sécurité et de confiance de ces applications ».

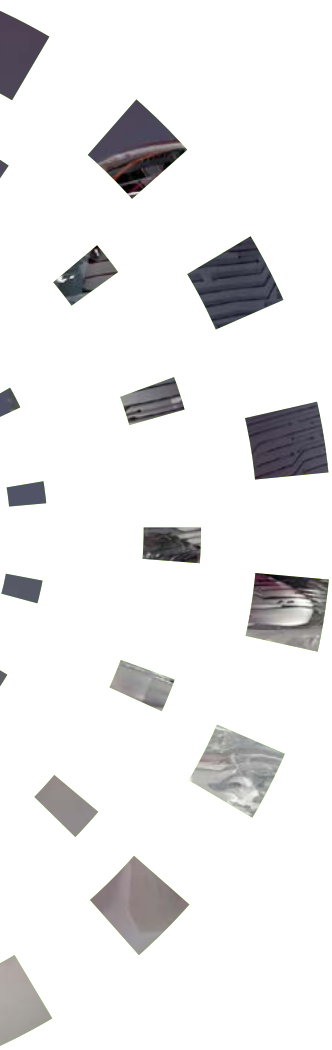
Benoît Tabaka, confirme la menace et observe aujourd'hui un mésusage grandissant de l'IA de la part de certains pays. « Ces pays, je ne vais pas les citer, vous les connaissez tous : nous pouvons évoquer les attaques informatiques du fameux groupe APT29, en lien avec la Russie, qui ont visé la CDU en Allemagne et bien d'autres cibles.

Bien évidemment, on peut s'attendre à d'autres attaques de cette nature visant des partis politiques. Nous avons pu détecter des *deepfakes* mettant en scène un président de la République dans une allocution contre des sans-abris ou des candidats aux élections européennes en train de jouer faussement à des jeux vidéo. Pour tout simplement déstabiliser. Les mésusages sont susceptibles d'apparaître, notamment au moment des élections. C'est notamment à cette occasion que nous devons jouer collectivement. Il faut mettre en place un certain nombre de garde-fous. Du côté des plateformes qui ont recours à l'IA mais qui sont aussi moteurs de recherche ou diffuseurs de produits vidéo – je pense à YouTube – nous devons mettre en place des labellisations, de l'information du grand public, mais aussi des outils de détection. Les deux cas de *deepfake* évoqués, ce sont nos outils qui les ont détectés sur la plateforme et nous ont permis de vérifier qu'effectivement, nous étions en présence de montages. C'est vrai qu'on a aujourd'hui un besoin de jouer collectif. Nous n'en sommes qu'au début, un peu comme lorsqu'on posait les tout premiers rails du chemin de fer, que l'on construisait les toutes premières lignes électriques des centrales. Il nous faut imaginer tous les usages qu'on va connaître en matière d'intelligence artificielle, d'intelligences artificielles.

Tout ne fait que commencer. C'est pour cela que les enjeux de régulation doivent être perçus dès maintenant, que les différentes étapes doivent être conçues dès maintenant, et que le rôle de tous les acteurs, et notamment en termes de cybersécurité doit être précisé. Sinon, nous serons face à de graves problèmes ».

De cet échange sur la cybersécurité on retiendra que le titre du 1er FIC (2007) était prémonitoire : « La cybercriminalité, criminalité du XXIe siècle ». Il est nécessaire de réglementer les usages, mais il faut aussi contrer les mésusages, face à un adversaire qui exploite pleinement les opportunités de l'IA. Lutter contre la cybercriminalité, phénomène planétaire, appelle une réponse planétaire. « Jouer collectif », c'est développer la coopération internationale, unir les efforts en matière d'information, de formation, de partage des savoir-faire, le tout dans une cadre légal qui respecte l'État de droit mais qui n'est pas non plus naïf. Les instruments existent, comme la Convention de Budapest du 23 novembre 2001, enrichie par ses deux protocoles additionnels. Mais les États prédateurs ne l'ont pas ratifiée. La lutte contre les prédateurs rappelle le duel du canon et de la cuirasse. Celui qui bénéficie de la rupture technologique prend le dessus.

agora
INCYBER FORUM



RETROUVEZ NOS DERNIÈRES ACTUALITÉS ET
NOS PROCHAINS ÉVÉNEMENTS SUR :

europe.forum-incyber.com