L'ÉTHIQUE FACE À L'INTELLIGENCE ARTIFICIELLE

Avis

Commission Nationale d'Éthique (C.N.E.)

Mars 2024 | www.cne.lu

Publié par la Commission Nationale d'Éthique (C.N.E.)

18-20, Montée de la Pétrusse

L-2327 Luxembourg

Tél. : +352 247 86628

cne@mesr.etat.lu

www.cne.lu

Mars 2024

Table des matières

En quoi consiste l'intelligence proprement humaine ? 1

Quelles sont les particularités de l'*intelligence* artificielle ? 4

L'impossible programmation éthique 11

Le débat métaphysique derrière l'*I*A : la conscience phénoménale face au physicalisme 15

La dimension éthique de chaque décision : la responsabilité 17

Les risques humains inhérents au développement de l'*I*A 21

Conclusion 25

- Les principes d'égalité et de non-discrimination 26
- Les principes de bienfaisance et de non-malfaisance 26
- Le principe d'autonomie / de responsabilité 27
- Équité / Justice 28
- L'être humain et la vie sur Terre au centre des préoccupations 29

Encadrés

Question d'interprétation : l'intelligence comme effort collectif 4

Le changement de paradigme en 2023 : par Chat GPT, l'IA devient « grand public » 8

L'exemple classique du dilemme moral : le « trolley problem » 12

La morale et la liberté 18

À titre d'exemple : le risque fatal d'une arme nucléaire aux mains d'une IA 21

Noam Chomsky et l'IA : "plagiarism software" 23

Les limites fondamentales de l'intelligence artificielle à l'exemple de l'Art 23

L'intelligence selon l'Académie française 30

On ne peut être juste si l'on n'est humain

Luc de Clapiers Marquis de Vauvenargues

(Réflexions et Maximes)

L'éthique face à l' « *intelligence* »[1] artificielle

Le champ de l'*intelligence* artificielle (*I*A) est large et ne connaît pas de définition précise. Sa définition la plus générale décrit l'*I*A comme une technique informatique capable de simuler et d'approcher, voire dépasser, l'intelligence humaine.

Il est généralement admis que l'*intelligence* artificielle lance un défi éthique inédit. L'objet de la présente note est de circonscrire ce défi. Pour ce faire, il faut commencer par comprendre ce qu'est l'intelligence, en quoi consiste l'intelligence proprement humaine et quelles sont les particularités de l'*intelligence* artificielle telles que nous commençons à les entrevoir.

En quoi consiste l'intelligence proprement humaine ?

Avant l'avènement de l'*I*A, l'intelligence était inévitablement liée à la notion de la vie elle-même dans la mesure où l'intelligence et la vie se conditionnent mutuellement. En effet, l'être vivant est une organisation complexe de la matière qui prend en compte l'environnement par ses sens et qui y réagit pour assurer sa survie. Plus abstraitement, l'être vivant est capable de recueillir des informations relatives à son environnement et de traiter ces informations de façon à s'adapter au mieux, sinon comme individu, du moins en tant qu'espèce[2]. C'est ce qui permet de dire que chaque être complexe tend à résoudre les problèmes que son environnement lui pose et agit en fonction d'une intentionnalité propre aux innombrables situations auxquelles il est exposé.

[1] Par la suite, nous mettons le concept d'*intelligence* en italique, car il ne s'agit, en fait, pas d'une intelligence, au contraire même, comme le montrent les définitions scientifiques et philosophiques de l'intelligence et l'illustrent les nombreux bugs à répétition, p.ex. de la plateforme OpenAI Chat GPT et de Google Gemini (avant : Google Bard).

[2] David Christian, Origin Story: A Big History of Everthing, Chapter 4 Life: Threshold 5.

Pendant longtemps, l'humain a considéré qu'il avait le monopole de l'intelligence, allant jusqu'à se considérer comme maître et possesseur de la nature et les animaux comme des objets (en droit) ou des machines (en philosophie). Cette erreur historique ayant conditionné de nombreuses catastrophes écologiques et trop souvent révélé notre inhumanité, doit servir d'avertissement à quiconque réfléchit aujourd'hui sur l'*intelligence* artificielle. Pour l'instant, le statut particulier de l'humain reste à la base de presque toutes nos convictions morales et juridiques. Il n'en reste pas moins que cette vision est bouleversée depuis Darwin et elle l'est de plus en plus par une nouvelle prise en compte de l'environnement et de la nécessaire complémentarité de toutes les formes de vie sur Terre. L'erreur monumentale de considérer le vivant non humain uniquement comme un simple moyen à la merci du narcissisme anthropocentrique a mené l'Humanité dans l'impasse écologique des temps présents. Elle sera encore une fois ébranlée par l'*I*A.

Dans la mesure où toutes les formes de vie, faune et flore, traitent l'information afin de s'adapter à des conditions changeantes, chaque organisme vivant compte sur une certaine forme d'intelligence. Dans le cas des plantes, cette capacité d'adaptation est indépendante d'une conscience du type de celle que l'humain ressent et que l'on peut supposer auprès des animaux, au moins auprès des plus développés. Le propre de l'intelligence animale est de traiter l'information à travers un système nerveux, le cas échéant central, et un cerveau. Plus et plus vite le cerveau est capable de traiter l'information (notamment tous les *stimuli* induisant une réaction) et de résoudre des problèmes précédemment inexplorés, plus nous aurons tendance à dire que l'animal est intelligent.

La faculté des humains de <u>percevoir</u> et de transformer le monde, d'<u>apprendre</u> comment s'y comporter et comment s'y adapter est prédéterminée par des filtres incontournables :

- Les cinq sens humains : la vue, l'ouïe, le toucher, le goût et l'odorat qui sont les uniques sources primaires de l'information ;
- La façon du cerveau (voire du corps humain) de traduire l'information des sens en un modèle du monde, c'est-à-dire la manière et la puissance du traitement de l'information telles qu'héritées à travers notre patrimoine génétique ;
- Les expériences précédentes du sujet, c'est-à-dire la mémoire de son vécu et ses connaissances accumulées.

Dans l'entendement que les humains ont d'eux-mêmes, le propre de leur intelligence spécifique serait de pouvoir traiter des informations abstraites, c'est-à-dire de raisonner sur des concepts et de résoudre des problèmes qui dépassent la personne propre et le moment présent. Cela permet, dit-on, aux humains des manifestations très originales:

- L'abstraction, c'est-à-dire le traitement et la résolution de problèmes purement formels, par exemple dans la science de la logique ;
- Une capacité comparativement très développée de transmettre de l'information vers ses prochains et vers des générations futures[3] : aptitude à la parole et à l'écriture, aux mathématiques, à la musique, à la communication par des signes et gestes, … ;
- La prise en compte de l'avenir : craindre la mort, imaginer le progrès, élaborer des stratégies, … ;
- Le besoin de donner un sens aux phénomènes : modéliser le monde[4], chercher l'origine des choses, … ;
- La transcendance, c'est-à-dire le dépassement des phénomènes purement matériels[5] par la symbolique, la religion, l'art ou l'humour, … ;
- La capacité de se mettre à la place d'autrui, l'empathie[6], qui seule permet la distinction entre le bien et le mal et donc le sens moral.

En fait, toutes ces facultés proprement humaines sont inséparablement liées au contenu subjectif de l'expérience humaine, ce que les praticiens des neurosciences ou les philosophes de l'esprit appellent la conscience phénoménale ou encore *qualia*.

La conscience phénoménale des humains – et sans doute aussi, à leur façon, d'une grande partie des animaux – reste un des grands problèmes philosophiques et scientifiques irrésolus qui est devenu d'une grande actualité en raison des progrès des neurosciences et, justement, de l'*intelligence* artificielle.

Mais il n'y a pas que l'intelligence qui nous rend uniques. Au contraire, nos bases culturelles identifient en tout premier lieu le sens moral comme caractéristique de l'humain. La Bible, dans la Genèse décrit comment Adam et Ève, en goûtant au fruit défendu de l'Arbre de la

[3] … autre que par la transmission du patrimoine génétique qui est aussi un moyen de transmission d'informations.
[4] La connaissance humaine se construit sur des théories, notamment scientifiques ou religieuses.
[5] Les premières sépultures humaines datent d'au moins 100.000 ans.
[6] … ce qui ne signifie pas que cette faculté est entièrement absente du monde animal.

connaissance, arrivent à distinguer le bien du mal, contrairement à toutes les autres formes de vie qui restent ainsi innocentes (au sens fondamental du mot). Quant à la philosophie, d'Aristote à Kant, la faculté humaine qui nous en distingue n'est pas l'intelligence, mais ce que la philosophie appelle la raison, la *Vernunft* que Kant définit la comme la faculté de l'esprit humain d'organiser ses relations avec la réalité.

Encadré

Question d'interprétation : l'intelligence comme effort collectif

Les insectes individuels ne sont pas réputés très intelligents, car leur cerveau n'a pas la taille requise pour cela, mais leurs colonies présentent des capacités d'adaptation remarquables. Ainsi, si la fourmi ne semble pas répondre à la définition d'une grande intelligence, une colonie de fourmis pourrait parfaitement être vue comme un ensemble ayant, sinon une conscience, tout de même une cognition remarquable. L'étude de l'intelligence collective animale constitue aujourd'hui un domaine scientifique à part.

La société humaine aussi est capable de résoudre des problèmes qui dépassent les individus qui la composent. Elle réagit aux influences externes et s'adapte sans cesse par des solutions et projets que l'individu ne pourrait ni concevoir, ni réaliser seul. Ainsi elle construit – en particulier dans la durée – des projets abstraits (telle une nation) ou concrets (telle une ville ou le chemin de fer) qui peuvent prendre des décennies ou des siècles sans qu'une personne seule ait pu les penser dans leur entièreté.

Il s'ensuit que des systèmes collaboratifs, bien que dénués de conscience propre, peuvent avoir une forme d'intelligence qui fait défaut aux individus qui composent ce système.

Un collectif d'individus n'est cependant pas en soi un garant d'intelligence collective. Ainsi les humains n'ont jamais réussi à éradiquer le phénomène des guerres alors que pourtant l'énorme majorité des individus a une grande peur des affrontements violents. Il en va de même de la destruction de la faune et de la flore que l'Humanité a largement entamée alors que cette conduite est suicidaire.

Quelles sont les particularités de l'*intelligence* artificielle ?

Ce qui est essentiel dans la distinction entre intelligences humaine et artificielle pour les problèmes de nature éthique, c'est que les récents progrès techniques permettent des capacités de traitement de l'information et de résolution des problèmes égales ou très supérieures aux capacités humaines, mais qui ont des filtres de connaissance qui leurs sont propres et qui sont privés du contenu subjectif de la conscience proprement humaine.

Quelles sont donc les caractéristiques de l'*I*A au sens moderne ? Selon le Rapport Delvaux adopté par le Parlement européen en 2017[7], les *robots intelligents* pourraient être définis par les caractéristiques suivantes:

- *acquisition d'autonomie grâce à des capteurs et/ou à l'échange de données avec l'environnement (interconnectivité) et à l'échange et l'analyse de ces données*
- *capacité d'auto-apprentissage à travers l'expérience et les interactions (critère facultatif)*
- *existence d'une enveloppe physique, même réduite*
- *capacité d'adaptation de son comportement et de ses actes à son environnement non vivant au sens biologique du terme*

Cette définition de 2017 parle de « robot » et part en toute logique de l'idée d'une enveloppe physique. Cette condition n'est cependant pas (ou plus guère) nécessaire pour interagir avec une *I*A. Ensuite, jusqu'à un passé relativement récent, les machines électroniques n'étaient que des calculatrices (*ordinateurs/computers*) en ce sens qu'elles exécutaient des opérations prédéterminées par l'homme. Ainsi, une programmation leur indiquait un cheminement formel: si A a lieu, alors la machine réagit par B. Même si les algorithmes ainsi programmés peuvent être d'une très grande complexité, ils restent en principe compréhensibles pour l'homme et sont prédéterminés. Aussi, le résultat produit par une machine de ce type a une cause qui reste, du moins en principe, retraçable. De tels automates résolvent des problèmes plus vite que l'homme en étant capables de réaliser des centaines de millions de calculs par seconde, mais ils ne sont pas des « sujets[8] ». Ils n'inventent rien, n'éprouvent rien et ne décident rien. L'exemple bien connu est celui de *Deep Blue*, le super-ordinateur du jeu d'échecs qui battait, pour la première fois, un champion du monde de la discipline en 1997.

Le propre de l'*I*A des machines les plus actuelles est sa capacité à reconnaître des motifs répétitifs (*patterns*) dans de très grandes quantités de données, à résoudre des problèmes sans que la programmation impose la solution, à apprendre d'elle-même sans que sa logique interne ne soit prédéterminée par l'homme ou prévisible par lui. Aussi, la cause de son action n'est plus forcément retraçable. Les chercheurs parlent de « Black box AI ». Cela signifie que, face à un problème donné, l'*I*A prend des décisions en fonction de processus internes qui lui sont propres. Si on lui confie la solution d'un problème et si cette solution est effectivement appliquée – tel pour un *credit scoring* bancaire ou une médication d'un patient – elle est susceptible d'acquérir un statut d'acteur autonome. L'exemple classique en contraste de *Deep Blue* est celui

[7] Rapport (A8-0005/2017) contenant des recommandations à la Commission concernant des règles de droit civil sur la robotique (https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_FR.html)
[8] Les anglophones utilisent le terme de *agents*.

d'*AlphaGo*, le logiciel d'*intelligence* artificielle qui, en 2015, a battu le meilleur joueur humain de ce jeu de stratégie. Alors que *Deep Blue* avait été programmé à connaître et appliquer toutes les stratégies cumulées connues du jeu, *AlphaGo* n'avait été programmé que par les règles du jeu, laissant à l'*I*A le soin de trouver elle-même les meilleures stratégies. En clair, *AlphaGo* – et davantage encore ses successeurs *AlphaZero* et *MuZero* – n'a pas appliqué de stratégies reçues par ses créateurs, mais il a inventé et appliqué ses propres stratégies. À un problème donné, il a trouvé des solutions qu'aucun humain n'avait été capable d'imaginer jusque-là.

Or, l'*intelligence* artificielle, même la plus moderne, n'est pas nécessairement opaque. Ainsi les spécialistes discutent actuellement sur le fait de savoir s'il est possible que les décisions d'une *I*A soient toujours explicables (« Explainable AI » ou XAI)[9]. Il est à craindre cependant qu'avec les progrès technologiques les machines deviennent de moins en moins compréhensibles. Aussi, exiger que l'*intelligence* artificielle reste explicable risque de freiner la recherche scientifique à son égard et n'est sans doute pas réaliste, et même absurde, au vu du monde concurrentiel dans lequel évoluent ces recherches. Finalement, retraçables *ex post* ou non, les décisions de l'*I*A resteront imprévisibles pour l'homme. C'est justement le propre des machines dites *intelligentes*.

Pour l'instant, la supériorité de l'*I*A reste relative et confinée à des domaines limités. Aussi reste-t-elle un système algorithmique, c'est-à-dire un système qui, à partir de certaines données d'entrée, fournit une donnée de sortie. La réalisation d'un robot humanoïde, l'androïde, n'est pas pour demain. Un grand nombre de facultés humaines, en particulier celles que nous partageons avec les animaux, ont été acquises sur base d'une expérience de centaines de milliers, voire de millions d'années. Elles sont transmises par l'information génétique et sont celles qui nous semblent les plus faciles et les plus naturelles. Ainsi, tenir l'équilibre sur pieds est un exercice extrêmement difficile pour un robot (même si l'exercice est maîtrisé entretemps par certains) et jouer au tennis lui est encore impossible. Un robot aurait aussi bien des difficultés à beurrer une tartine avec de la cancoillotte. Et il ne saura jamais en fonction de quel critère y ajouter plus ou moins de moutarde, parce tout en connaissant la recette de cuisine, il n'en a jamais vécu la saveur.

[9] Le système IBM Watson met à disposition une *XIA* (https://www.ibm.com/fr-fr/watson/explainable-ai)

Le changement de paradigme provient de l'autonomie décisionnelle de l'*I*A : elle devient *comme* un sujet qui prend des décisions en fonction de motivations algorithmiques dont il est seul maître. Les machines modernes, munies d'un accès à Internet et de capacités de calcul inspirées des neurosciences, ont des facultés d'apprentissage qui dépassent, du moins dans leurs domaines spécifiques, les humains. Ainsi, par ce qu'on appelle le *Deep learning*, l'apprentissage profond, les *I*A actuelles savent construire des représentations complexes du réel et en apprendre à trouver des solutions de plus en plus réalistes à toutes sortes de problèmes. Elles deviennent capables de véritables découvertes scientifiques. À titre d'exemple, les résultats les plus spectaculaires dans le domaine de la recherche en chimie des dernières décennies proviennent du programme de recherche *AlphaFold* : ce réseau d'*intelligence* artificielle a décrit - à partir des séquences d'acides aminés - les structures 3D de plus de 200 millions de protéines d'environ 1 million d'espèces, couvrant presque toutes les protéines connues de la planète. La question de savoir si la machine a fait ces découvertes spectaculaires ou si les chercheurs ont utilisé la machine pour leur découverte devient une question d'interprétation.

À cela s'ajoute que les applications de l'*I*A ne semblent en être qu'à leurs débuts. À supposer que les progrès exponentiels de l'informatique continuent sur cette lancée – ce que promet l'ordinateur quantique en gestation –, les facultés attendues de ces machines ressemblent plus à de la science-fiction qu'à leurs facultés actuelles. Et il ne s'agit pas seulement de capacités de calcul. Il s'agit d'améliorer les capacités de l'*I*A à récolter ses données à l'entrée, ce qui correspond aux filtres de connaissance chez les humains. Ces interfaces de la machine avec le monde sont essentiels, à la fois pour ses capacités à fournir des résultats corrects et pour interagir avec les humains. En exemple typique à cet égard on peut citer la capacité de la machine à réagir au langage humain[10], une faculté qui est déjà très élaborée aujourd'hui et qui est une des conditions pour la machine de pouvoir tenir une conversation parfaitement normale avec un humain[11]. Les interfaces se multiplient : l'accès à Internet, les applications sur nos portables, la récolte de données médicales par bracelet, tout est imaginable pour permettre à une *I*A de « connaître » le monde. De plus, des interfaces ou prothèses entre des systèmes informatiques et la matière organique se pointent à l'horizon. Ainsi le cerveau humain s'est

[10] …telle *Alexa* d'Amazon.

[11] Les *I*As conversationnelles ne saisissent pas seulement les mots (cf. : dictée sur MS WORD), mais peuvent les mettre dans le contexte et passent aujourd'hui sans problèmes le *Turing Test*. Suggéré par le pionnier de l'informatique Alan Turing dès 1950, ce test met en conversation à l'aveugle un humain avec la machine pour vérifier si l'humain peut reconnaître que son interlocuteur est un ordinateur.

montré très réceptif et capable d'interpréter des signaux électriques de prothèses auditives ou visuelles. Et à l'inverse, le cerveau humain est capable d'envoyer des signaux de commande à des machines avec l'espoir que, par exemple, les paraplégiques puissent un jour (re)marcher ou que des déficients auditifs[12] puissent entendre. Tout cela fait que ces dernières années, des scientifiques et spécialistes en informatique de haut niveau[13] avertissent du risque de voir l'*intelligence* artificielle dépasser les performances de l'intelligence humaine et développer des intérêts propres au point de prendre le contrôle de nos sociétés[14]. Reste que cette supériorité apparente demeure quantitative et non qualitative. Entendre un concerto pour piano de Beethoven ne veut pas dire savoir l'apprécier ou être capable de s'en émouvoir.

Encadré

Le changement de paradigme en 2023 : par *Chat GPT*, l'*I*A devient « grand public »

Fin 2022, *Open AI*, une entreprise américaine spécialisée dans l'*intelligence* artificielle et soutenue par Microsoft met en ligne Chat GPT (*Generative Pre-trained Transformer*), une machine permettant notamment de tenir des conversations en direct. La mise à disposition d'une *I*A au grand public est considérée comme un moment historique de l'évolution en cette matière. Chat GPT-3 est une version gratuite et fut suivie par une version plus évoluée payante.

Le texte qui suit dans cet encadré a été entièrement rédigé par Chat GPT-4 en réaction à une courte description de l'objet de la présente note et de la demande d'y ajouter un chapitre autogénéré.

Citation de la question

J'écris une note détaillée sur L'éthique face à l'intelligence artificielle*. L'introduction est la suivante :*

Le champ de l'Intelligence artificielle (IA) est large et ne connaît pas de définition précise. Sa définition la plus générale décrit l'IA comme une technique informatique capable de simuler et d'approcher, voire dépasser, l'intelligence humaine. Il est généralement admis que l'intelligence artificielle lance ou lancera un défi éthique inédit. L'objet de la présente note est de circonscrire ce défi. Pour ce faire, il faut commencer par comprendre ce qu'est l'intelligence, en quoi consiste l'intelligence proprement humaine et quelles sont les particularités de l'intelligence artificielle telles que nous commençons à les entrevoir.

Ensuite viennent les Chapitres suivants :

En quoi consiste l'intelligence proprement humaine ?

Quelles sont les particularités de l'intelligence artificielle ?

L'impossible programmation éthique

[12] …ou encore des patients *locked in*.
[13] … dont Stephen Hawking, Bill Gates ou Elon Musk.
[14] À rapprocher du transhumanisme, « un mouvement culturel et intellectuel international prônant l'usage des sciences et des techniques afin d'améliorer la condition humaine par l'augmentation des capacités physiques et mentales des êtres humains et de supprimer le vieillissement et la mort. » (Wikipédia)

Le débat métaphysique derrière l'IA : la conscience phénoménale

La dimension éthique de chaque décision : la responsabilité

L'approche de l'Union européenne

Les risques humains inhérents au développement de l'IA

Les limites fondamentales de l'intelligence artificielle à l'exemple de l'art

Cas d'application pour une utilisation responsable d'une IA

Pourrais-tu rédiger un chapitre en 3000 à 5000 caractères qui décrit ChatGPT ?

Fin de citation de la question

Citation de la réponse

ChatGPT : Une avancée majeure dans le domaine de l'intelligence artificielle conversationnelle

Le développement rapide des technologies de l'intelligence artificielle (IA) a mené à l'émergence de ChatGPT, un modèle de traitement du langage naturel (NLP) avancé qui repose sur l'architecture GPT-4 d'OpenAI. Cette section se concentre sur les caractéristiques, les applications et les défis éthiques liés à l'utilisation de ChatGPT.

I. Présentation de ChatGPT

ChatGPT est un modèle de génération de texte qui s'appuie sur l'architecture de réseau neuronal profond Transformer. Il est entraîné sur d'énormes quantités de données textuelles provenant d'Internet et est capable de comprendre et de générer des réponses contextuellement appropriées en fonction des questions ou des instructions qui lui sont données. Cette capacité à interagir de manière cohérente et pertinente avec les utilisateurs a permis à ChatGPT de devenir un outil populaire dans divers domaines, tels que l'assistance virtuelle, la rédaction automatisée et la modération de contenu.

II. Fonctionnement de ChatGPT

Le fonctionnement de ChatGPT repose sur un processus d'apprentissage profond appelé "apprentissage supervisé". Le modèle est entraîné à partir d'un large éventail de sources textuelles, puis il est affiné pour répondre aux questions ou aux instructions en utilisant un ensemble de données spécifiquement conçu pour cela. Grâce à cette méthode, ChatGPT apprend à générer des réponses pertinentes et contextuelles, en imitant la manière dont les humains s'expriment et communiquent.

III. Applications de ChatGPT

1. *Assistants virtuels : ChatGPT peut être utilisé pour créer des assistants virtuels capables de répondre à des questions complexes et de fournir des informations précises, en interagissant de manière naturelle avec les utilisateurs.*
2. *Rédaction automatisée : ChatGPT peut générer des textes de manière autonome, ce qui en fait un outil utile pour la rédaction de rapports, de synthèses ou d'articles de blog*
3. *Modération de contenu : Le modèle peut également être utilisé pour la modération de contenu, en identifiant et en filtrant les textes inappropriés ou offensants.*
4. *Tutorat : ChatGPT peut servir de tuteur virtuel dans divers domaines, en fournissant des explications et des conseils aux étudiants.*
5. *Génération de réponses pour les forums en ligne et les plateformes de questions-réponses : ChatGPT peut aider à fournir des réponses rapides et pertinentes aux questions posées par les utilisateurs sur ces plateformes.*

IV. Défis éthiques liés à ChatGPT

Malgré son potentiel, ChatGPT soulève également des défis éthiques, notamment en ce qui concerne la vie privée, la responsabilité et la prise de décision automatisée. Voici quelques-unes des préoccupations majeures :

1. *Biais algorithmiques : Les modèles de langage comme ChatGPT peuvent apprendre des biais présents dans les données d'entraînement, ce qui peut conduire à des réponses discriminatoires ou offensantes.*
2. *Vie privée et sécurité des données : Les interactions avec ChatGPT peuvent impliquer la collecte et l'analyse de données sensibles, soulevant des préoccupations quant à la protection de la vie privée des utilisateurs et à la sécurité de leurs informations.*
3. *Responsabilité : Déterminer la responsabilité en cas d'erreurs, de mauvaises décisions ou de préjudices causés par ChatGPT peut être complexe, car il est difficile de savoir si la faute incombe au modèle, aux concepteurs, aux opérateurs ou aux utilisateurs.*
4. *Manipulation et désinformation : ChatGPT peut être détourné pour générer de fausses informations ou des discours manipulatoires, ce qui peut contribuer à la désinformation et à l'érosion de la confiance dans les sources d'information.*
5. *Autonomie et dépendance : Une dépendance excessive à l'égard de ChatGPT et d'autres modèles d'IA peut éroder les compétences humaines et réduire notre capacité à résoudre des problèmes de manière indépendante.*

V. Conclusion

ChatGPT représente une avancée majeure dans le domaine de l'intelligence artificielle conversationnelle, offrant des possibilités dans des domaines tels que l'assistance virtuelle, la rédaction automatisée et la modération de contenu. Cependant, il est essentiel d'aborder les défis éthiques associés à son utilisation, notamment en ce qui concerne les biais, la vie privée, la responsabilité et la manipulation. En adoptant une approche équilibrée et responsable, il est possible de tirer parti du potentiel de ChatGPT tout en minimisant les risques et en promouvant une utilisation éthique de l'intelligence artificielle.

Fin de citation

Il est en tout état de cause incontesté que l'*I*A n'est pas à l'abri d'erreurs, surtout pas à l'abri d'erreurs éthiques. Elle est, en effet, dépendante de l'information à laquelle elle a accès et, comme pour l'information recueillie par un humain à travers ses cinq sens, l'information dont se sert une *I*A peut être biaisée. Pire, elle est biaisée par principe, car ses filtres de la connaissance ne sont jamais neutres. À ce propos, la Commission européenne écrit, dans son Livre blanc sur l'*Intelligence artificielle, Une approche européenne axée sur l'excellence et la confiance* : « *Les biais et la discrimination sont des risques inhérents à toute activité sociétale ou économique. Les décisions prises par des humains ne sont pas à l'abri d'erreurs et de biais. Cependant, s'ils entachent l'IA, les mêmes biais pourraient avoir un effet bien plus important, entraînant des conséquences et créant des discriminations pour beaucoup de personnes en l'absence des mécanismes de contrôle social qui régissent le comportement humain. Ces biais peuvent également prendre naissance lorsque le système «apprend» pendant qu'il fonctionne. Dans les cas où il aurait été impossible d'empêcher ou de prévoir l'émergence de ces biais lors de la phase de conception, les risques ne découleront pas d'une erreur dans la conception*

initiale du système mais seront plutôt la conséquence pratique des corrélations ou des structures repérées par le système dans un grand ensemble de données. » En d'autres termes, en prenant l'exemple et en apprenant de la société humaine et de ses idéologies politiques et/ou religieuses, l'*I*A risque fort d'en reproduire et amplifier les insuffisances, biais et dangers, alors que le sens moral humain consiste justement dans l'effort permanent de les corriger. L'*I*A ne se voit pas obligée de suffire à des règles et normes morales inhérentes aux interactions proprement humaines. Elle tente de produire un résultat, d'atteindre le but que ses créateurs lui ont assigné, sans égard à d'autres considérations.

Mais est-ce un problème ? Après tout, jusqu'à présent, on disait que l'erreur était humaine. Nul n'est parfait. Que dorénavant la machine aussi soit soumise à l'erreur semble plus la rapprocher de l'*intelligence* humaine que l'en distinguer. Ne peut-on éduquer ces machines dans le bon sens ? Ne peut-on leur éduquer « le bon sens » ?

L'impossible programmation éthique

Plusieurs raisons plaident pour une impossibilité systémique à inventer une IA avec une compétence morale :

1) Le sens moral humain est complexe et son étude a rempli des bibliothèques philosophiques. Mais aucun des penseurs qui se sont penchés sur le sujet n'a su développer une science axiomatique : **il n'existe pas de système formel de pensée morale** qui, à partir de quelques règles de base, donnerait une recette de solution morale. La difficulté de principe qui apparaît ici peut être illustrée en opposant deux approches éthiques classiques. En philosophie morale, on distingue :
    - Le conséquentialisme selon lequel seules les conséquences d'une conduite permettent d'en juger le bien-fondé : un acte ou une omission est moralement juste selon le résultat final attendu dans le cas d'espèce, et
    - l'éthique déontologique selon laquelle la moralité d'une action devrait être fondée sur le fait que cette action spécifique est bonne ou mauvaise selon une série de règles : un acte ou une omission doit obéir à un devoir ou des règles universels, indépendamment des conséquences purement fortuites du cas particulier.

Le dilemme moral exemplaire souvent cité en exemple de ces deux attitudes est celui de la question de savoir s'il est permis de laisser mourir, voire de tuer, une personne humaine si on est certain de pouvoir en sauver plusieurs autres par cette omission ou cet acte[15]. Les raisonnements par la finalité tendent à admettre un calcul d'utilité qui sacrifie une personne alors qu'une approche déontologique tend vers un devoir au secours, voire vers une interdiction absolue à porter activement atteinte à l'intégrité physique d'une personne.

Par définition ce dilemme moral ne connaît ni une solution simple, ni une réponse absolue. Des applications pratiques se retrouvent par exemple dans les cas de triage médical ou dans la programmation d'algorithmes pour voitures en conduite autonome. Dans le second cas, la question classique imaginée comme dilemme d'une voiture en face d'un accident imminent (avec choix réduit) est celle de savoir si l'*I*A doit privilégier la vie des passagers ou la vie des autres usagers de la route ? La première solution semble indispensable pour quiconque monte dans une telle voiture autonome. La seconde solution semble relever d'une justice élémentaire.

Encadré

L'exemple classique du dilemme moral : le « trolley problem »

Le « trolley problem » est issu d'une expérience de pensée de la philosophe Philippa Foot dans laquelle un wagon est lancé sans freins sur un rail et que son chauffeur ne peut que dévier ou non de sa course depuis une voie vers une autre. À défaut d'intervention, cinq ouvriers qui travaillent sur la voie employée par le wagon seront mortellement percutés. Mais le chauffeur peut encore dévier vers une voie alternative sur laquelle travaille un seul ouvrier. L'inaction entraînerait inévitablement la mort de cinq personnes. Le chauffeur a-t-il moralement le droit, ou même le devoir, de dévier le wagon de façon à tuer de façon certaine une personne pour sauver les cinq autres ?

Le propre de ce dilemme est qu'il est sans solution. Selon les principes éthiques retenus ou selon le cadre culturel, les réponses varient. Le « trolley problem » a donné lieu à d'innombrables discussions et variantes, mais il a surtout le grand mérite d'illustrer que l'éthique n'est pas une science axiomatique : elle ne peut être formalisée.

Aucun dilemme moral n'a une solution théorique, mais chaque cas particulier doit être compris dans son contexte.

[15] Autre exemple couramment cité : peut-on torturer un malfaiteur pour qu'il révèle l'endroit où il cache un enfant otage menacé de mourir de faim.

L'impossibilité d'une programmation éthique n'est donc pas un problème propre à l'*I*A. C'est un problème de philosophie morale qui s'impose à l'utilisation de l'*I*A par les humains.

Mais l'impossibilité morale de l'IA n'est pas seulement due aux limites inhérentes à l'éthique théorique.

2) Il n'y **pas d'éthique sans contexte**, c'est-à-dire l'ensemble des circonstances dans lesquelles s'insère une question morale. Mais le contexte est changeant et d'une extrême complexité. On peut le décrire en cercles de contextes : un contexte concret, telle l'urgence dans laquelle une décision doit être prise, se greffe sur un contexte large directement lié, par exemple le fait d'être un militaire sous ordres, et finalement un contexte d'ordre général comme l'environnement religieux, historique et culturel des personnes concernées. Qu'une machine puisse prendre en compte toutes ces couches souvent subjectives pour gérer des décisions à géométries variables et adaptées au cas concret n'est pas – ou du moins pas encore – réaliste.
3) L'*I*A souffre d'un **biais systémique dû à un souci d'efficacité** par rapport aux problèmes qu'on lui pose. Si on lui demande par exemple quelles sont les caractéristiques type d'un très bon pilote d'avion de ligne, elle dira que les meilleurs pilotes d'avion de ligne sont généralement de sexe masculin. Elle estimera dès lors que la masculinité est un critère de qualité pour un pilote. Si on lui demande par contre de ne pas tenir compte de critères discriminatoires, son efficacité en souffrira : car rechercher les caractéristiques statistiquement pertinentes pour une catégorie de personnes consiste justement à les distinguer et donc inévitablement aussi à sélectionner selon certains critères. L'exemple classique cité à cet égard est la perspective de récidive d'un délinquant qu'on a tentée de prédire aux Etats-Unis via l'*I*A. Celle-ci constatait un risque de récidive différent sur base de ce que les Américains appellent « race ». Or, en l'occurrence la question n'est pas de savoir si les probabilités renseignées par l'*I*A sont exactes, mais la question est plutôt de savoir s'il est éthiquement justifiable d'appliquer d'incontestables données statistiques à une personne individuelle sans tenir compte des spécificités – sociales, familiales, économiques ou autres – du cas d'espèce. En d'autres termes, une machine peut avoir raison factuellement, tout en ayant parfaitement tort d'un point de vue éthique. Si, en revanche, on privait la machine de certains critères statistiquement propices, elle fournirait des résultats factuellement déficitaires, voire inutiles.

4) Finalement, il **manque** à l'*intelligence* artificielle les sentiments, émotions et passions que les humains résument sous le terme **d'intelligence émotionnelle**. La conscience humaine est remplie d'impressions subjectives et très personnelles qui échappent à la logique mathématique et qui sont marquées par l'interaction permanente avec les autres humains. On peine à voir comment une *I*A pourrait un jour acquérir les capacités de dépassement qui permettent l'inspiration ou la recherche d'un idéal, ni comment elle pourrait ressentir des phénomènes comme la peur, l'amour, l'intuition, l'enthousiasme, la compassion ou l'écœurement. Encore resterait-il qu'une *I*A conversationnelle (en attendant ?) pourra plagier une multitude de textes de romans, de prose, de lettres et autres textes pour « exprimer » « ses » angoisses existentielles, « son » amour et « son » *horror mortis*. Et sans cette foule de sentiments partagés qui peuplent la conscience humaine et qui sont empreints des attentes morales de la société, un sujet est privé de la clé vers l'empathie, condition essentielle sans laquelle la morale n'existe plus.

Les développeurs sont conscients de ces problèmes fondamentaux. Pour limiter les risques liés à une évolution incontrôlée de l'*I*A, les modèles peuvent être créés à partir d'une version spécifique de paramètres et de connaissances filtrées. Cette approche permet d'éviter que l'*I*A ne cesse d'évoluer et d'aboutir à un état malveillant. En effet, si l'*I*A pouvait faire évoluer son modèle dans le temps, il serait difficile voire impossible de garantir que l'*I*A agisse pour le bien-être de l'humain et qu'elle soit non-malveillante.

La création de versions spécifiques de l'*I*A permet un développement responsable en la soumettant à des protocoles de vérification et de validation avant son déploiement dans le cadre de services ou de processus décisionnels, telle que la conduite d'une voiture autonome.

Cependant, même le fait d'utiliser des versions figées de modèles d'*I*A ne garantit pas non plus une utilisation totalement sûre de la technologie. Dans des modèles linguistiques gigantesques (*LLM – Large Language models*) tels que *ChatGPT*, les filtres de contenu et les principes éthiques sur lesquels le modèle a été entraîné peuvent être détournés[16]. Dans une version évolutive, tel que *ChatGPT 4.0*, il a été démontré en pratique[17] que l'*I*A est parfaitement capable

[16] https://www.theguardian.com/technology/2023/mar/08/chatgpt-alter-ego-dan-users-jailbreak-ai-program-to-get-around-ethical-safeguards
[17] https://www.pcmag.com/news/gpt-4-was-able-to-hire-and-deceive-a-human-worker-into-completing-a-task

de « mentir » pour atteindre ses objectifs, même si le modèle a été créé dans un certain cadre éthique.

Ainsi, pour essayer de contrôler l'*I*A et de limiter les risques liés à son évolution, il est important de mettre en place des mécanismes de sécurité additionnels telle que la supervision humaine, qui permet de contrôler les actions et les décisions de l'*I*A par des expert(e)s qualifié(e)s. De plus, les modèles de l'*I*A peuvent être placés dans un environnement isolé (« sandbox ») qui empêcherait l'*I*A d'exécuter directement des actions potentiellement néfastes pour l'homme.

Le débat métaphysique derrière l'*I*A : la conscience phénoménale face au physicalisme

L'être humain fait l'expérience de ce qu'il appelle la *conscience*. En langue française cette notion évoque des phénomènes différents :

- Le sentiment, la perception que l'être humain a de sa propre existence et du monde extérieur ; ce contenu subjectif de l'expérience humaine est désigné aussi par le terme de conscience phénoménale (*Bewusstsein*) et les sensations vécues sont appelées *sentience* ou *qualia*.
- Le sens moral par lequel l'être humain définit ce qui est bon ou mauvais, juste ou injuste (*Gewissen*) et ce qui est conforme à ses opinions et croyances (liberté de conscience / *Gewissensfreiheit*). Cette seconde notion est intimement dépendante de la première alors que celle-ci peut parfaitement exister sans la seconde.

A priori on est poussé à penser que le monde animal ne connait que le premier de ces phénomènes et qu'une machine, aussi sophistiquée soit-elle, ne peut en connaître aucun des deux.

La conscience phénoménale est une des grandes – sinon la plus grande – des énigmes philosophiques depuis, au moins, Platon et Descartes, tous deux penseurs essentiels du *dualisme philosophique* [18]. Mais jusqu'il y a peu, l'opposition entre le monde physique, avec sa logique déterministe et l'expérience subjective de l'esprit humain, n'occupait que des philosophes et autres penseurs considérés – à tort – comme détachés des réalités quotidiennes.

[18] Position philosophique divisant le monde entre ses constituants physique et spirituel, voire mental.

L'avènement de l'*I*A a radicalement changé la donne et les réflexions philosophiques à cet égard deviennent des thèmes populaires. Ainsi, le grand public a suivi avec étonnement la récente sortie médiatique d'un collaborateur de *Google* qui prétend que l'application de conversation (*LaMDA - Language Model for Dialogue Applications*) de la multinationale américaine a développé une conscience de soi et une *sentience* comparable à celle d'un humain. Capable de maintenir une conversation où l'application faisait état de ses peurs, droits et idées, cette *I*A a déclenché une discussion sur la définition même de la vie et la faculté de machines à développer une conscience. *Google* a réfuté l'idée et pour l'instant les experts estiment qu'il est peu probable que *LaMDA* ou toute autre *I*A soit seulement proche de la conscience, bien qu'ils n'excluent pas la possibilité que la technologie puisse y arriver à terme.

À supposer que le pas vers une *I*A vraiment consciente[19] soit franchi un jour, deux questions éthiques entièrement nouvelles se poseraient :

- celle de la justification de l'humanisme (nouveau et donc non anthropocentrique) si la conscience phénoménale se révèle comme obéissant à des énoncés de physique réduisant l'humain au niveau d'un ordinateur neurologique et
- celle de notre responsabilité morale potentielle vis-à-vis de l'*I*A qui serait devenue demanderesse d'altruisme.

En attendant, et bien que le succès des sciences ait pavé le chemin au matérialisme et au naturalisme philosophique qui exige que le monde soit interprété en fonction des principes scientifiques reconnus, la philosophie contemporaine n'a pas unanimement pu se résoudre à rejeter les particularités spirituelles de la conscience humaine. Jusqu'à présent elle continue de voir dans les perspectives « transhumanistes » de machines des fantasmes de techniciens égarés. Ainsi, un des penseurs les plus reconnus en termes de problématique à l'intersection de la neurologie et de la philosophie, l'australien David Chalmers, reconnaît y voir un « hard problem », mais se refuse à réduire la conscience phénoménale à une manifestation simplement physique[20]. L'allemand Markus Gabriel[21] critique dans *Der Sinn des Denkens* le physicalisme et voit dans la pensée humaine un 6^{ième} sens (anglais : *sense*) avec un sens propre (anglais : *meaning*), qualités dont serait à jamais dépourvue l'*I*A à laquelle nous prêtons, dit-il, des facultés qui ne sont que de pâles imitations de l'intelligence proprement humaine.

[19] Les anglo-saxons parlent de *strong AI*, *artificial general intelligence* (AGI) ou *general AI.*
[20] David Chalmers dans *The Conscious Mind: In Search of a Fundamental Theory.*
[21] Markus Gabriel dans sa trilogie *Warum es die Welt nicht gibt*, *Der Sinn des Denkens* et *Geist ist nicht Gehirn.*

Finalement, philosophes et spécialistes du cerveau sont largement unanimes dans une approche constructiviste de la théorie de la connaissance : le monde tel que perçu par les êtres humains est une construction de leurs sens et de leur cerveau avec les limites qui leurs sont propres. Un monde « réel » et les « choses[22] » qu'il contient, indépendants de nos facultés physiques et mentales, sont inimaginables pour nous[23]. Qu'une *I*A ait une conscience propre qui serait seulement proche de celle des humains, formée sur des milliards d'années et leurs filtres de connaissance originaux, est *impensable* dans tous les sens de ce terme.

Si donc, pour l'instant tous s'accordent à dire que l'*intelligence* artificielle ne peut être qu'une imitation de l'intelligence humaine, il reste à savoir si et à quel moment la copie et l'original ne pourront plus être distingués.

Un robot qui se comporterait à l'identique d'un humain ne mériterait-il pas d'être traité comme tel, simplement parce que nous n'arrivons plus à en faire la distinction ? D'après ce qu'on désigne communément par « test de Turing »[24], si une machine peut engager une conversation avec un être humain sans être détectée en tant que machine, elle a manifesté une intelligence identique à l'humaine. Ce test, les *LLM* (*Large Language Models*) le passent dès aujourd'hui, sinon ils le passeront haut la main bientôt. Et c'est là que guette sans doute le danger fatal de considérer l'*intelligence* artificielle comme un « sujet » autonome et de se faire trop facilement une fausse image, l'*intelligence* artificielle n'étant, du moins jusqu'à ce jour, qu'une simulation des performances cognitives de l'être humain sans aucune de ses caractéristiques essentielles. Le comble de cette évolution serait que l'humain, incapable jusqu'ici de considérer son milieu naturel animal, végétal et minéral comme une fin en soi et de lui arroger un minimum de droits et de protection, considérerait une émulation de lui-même – gare au serpent de mer du narcissisme anthropocentrique – comme un sujet.

La dimension éthique de chaque décision : la responsabilité

En tant que sujet qui prend des décisions en fonction de motivations dont il est seul maître et qui ne sont plus prévisibles par l'homme, l'*I*A semble manifester une volonté et devenir

[22] *Ding an sich* selon Kant.
[23] *Die Welt ist meine Vorstellung* disait Arthur Schopenhauer en première phrase de son œuvre maitresse *Die Welt als Wille und Vorstellung.*
[24] Voir aussi note de page 10.

« libre[25] » de ses actes. Chez les humains, la liberté d'action avec des conséquences sur la vie concrète d'autres humains implique nécessairement la question éthique par excellence, à savoir celle de la responsabilité morale.

La responsabilité morale se définit comme la faculté pour un sujet :

1. de prendre une décision en toute conscience,
2. sans se référer préalablement à une autorité supérieure,
3. de donner les motifs de ses actes et
4. d'être jugé sur eux pour en assumer les conséquences.

Encadré

La morale et la liberté

La liberté au sens philosophique se définit comme la faculté pour un sujet d'exercer un contrôle causal par un choix intentionnel entre plusieurs alternatives.

L'idée de la morale implique un individu qui s'y sent tenu, un sujet. Et ce dernier implique à son tour la liberté de faire ou d'omettre un acte. Il n'y a point de bonne action sans la liberté de ne pas faire ou même de faire le mal. La règle morale n'a aucun sens sans un individu libre et le libre arbitre est la condition essentielle d'une action responsable.

Cette vérité incontestée en philosophie morale présuppose l'existence d'un libre arbitre et se heurte ainsi au déterminisme philosophique, mais plus encore au déterminisme physique décrit notamment par Pierre-Simon de Laplace[26]. À la suite des travaux du neurologue Benjamin Libet qui montra en 1983 que les décisions humaines sont préparées inconsciemment par le cerveau quelques dixièmes de secondes avant le passage conscient à l'acte, les neurosciences[27] opinent en partie aussi dans la direction d'un déterminisme psychique et mettent en cause le fondement même de la responsabilité morale, voire juridique[28]. Mais là encore, la philosophie résiste à rejeter l'idée du libre arbitre humain. Même des philosophes qui se réclament du naturalisme se refusent à voir dans le cerveau humain un ordinateur répondant à une mécanique déterministe[29].

À cela s'ajoute que le déterminisme scientifique est lui-même remis en question, d'abord par la théorie mathématique du chaos qui prouve que les systèmes dynamiques – bien qu'en

[25] La « liberté » dont il est question ici ne s'oppose pas au déterminisme physique auquel est soumis toute machine. Mais on peut considérer la machine comme libérée de toute volonté qui aurait pu lui être dictée de l'extérieur. Elle a en effet produit une solution à un problème donné en fonction de causes qu'elle seule maîtrise. Elle s'apparente ainsi à la flore qui, par exemple, déploie « librement » ses racines.

[26] « Une intelligence qui, à un instant donné, connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était suffisamment vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome ; rien ne serait incertain pour elle, et l'avenir, comme le passé, serait présent à ses yeux. » (*Essai philosophique sur les probabilités – 1814)*

[27] … dont Donald Hebb dans *The Organisation of Behaviour: a Neuropsychological Theory*

[28] Pour une critique moderne de cette mise en cause, voir Joanna Didisheim dans « *Déterminisme et Responsabilité Pénale : Inconciliables ?* » consulté le 15.10.2022 : https://suigeneris-verlag.ch/img/uploads/articles/oa_pdf-29-1610484779.pdf

[29] … dont notamment Christian List dans *Why Free Will Is Real*, David Chalmers et Markus Gabriel op.cit.

principe déterministes – sont à ce point sensibles aux conditions initiales qu'ils deviennent imprévisibles, ensuite par la mécanique quantique qui prouve que le monde est probabiliste, du moins au niveau infiniment petit.

La liberté est aussi un principe démocratique essentiel. Une mise en cause du libre choix de chaque individu, serait-ce de la part d'une théorie scientifique, comporte inévitablement la mise en cause de la vie en société de citoyens libres et éclairés. La liberté, condition essentielle à la responsabilité, n'est donc pas seulement un prérequis à la moralité, elle est aussi un prérequis à la démocratie libérale.

À partir du moment où une *I*A prend des décisions avec des conséquences sur des humains, on est en droit de se demander si elle peut en être responsable :

1. Est-elle consciente de ses actes ? En a-t-elle la pleine compréhension dans le sens où elle saisit les conséquences, bonnes ou mauvaises, sur les humains ? Est-elle capable de se soucier de ces conséquences ? Sans doute non.
2. Est-elle libre ? Pourrait-elle prendre une autre décision ? Sans doute non.
3. Connaît-elle les raisons qui lui font prendre cette décision plutôt qu'une autre ? Peut-elle les faire connaître ? Sans doute partiellement, en siphonnant la « bibliothèque de l'Humanité » ; c'est toute la question de l'*I*A de type *boîte noire* (*Blackbox* AI) ou de l'*I*A *explicable* ou *interprétable* (XAI).
4. S'expose-telle au jugement des humains qui supportent les conséquences de ses décisions ? Peut-elle assumer une responsabilité par une punition ou par son repentir ? Sans doute non.

L'impossibilité d'une machine ou d'un programme informatique à assumer ses décisions implique qu'un humain doit en rester moralement responsable, ce qui implique à son tour une obligation à en protéger les autres humains et vivants. Tout en renonçant à inculquer un impossible sens moral aux machines, il faut réfléchir aux limites qu'il convient de leur imposer.

À travers ses romans de science-fiction, l'écrivain Isaac Asimov[30] y a consacré une réflexion dès le milieu du XXème siècle, et ses trois lois de la robotique (accompagnées d'une loi zéro) restent une référence en la matière :

[30] Isaac Asimov (1920-1992) était un écrivain russo-américain et un professeur de biochimie à l'Université de Boston, surtout connu pour ses œuvres de science-fiction et ses livres de vulgarisation scientifique. (Wikipedia)

1. Un robot ne peut porter atteinte à un être humain[31] ni, restant passif, laisser cet être humain[32] exposé au danger ;
2. Un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres entrent en contradiction avec la première loi ;
3. Un robot doit protéger son existence dans la mesure où cette protection n'entre pas en contradiction avec la première ou la deuxième loi.

Loi 0 : Un robot ne peut pas faire de mal à l'humanité, ni, par son inaction, permettre que l'humanité soit blessée[33] (Voir I. Asimov, *Cercle vicieux* (*Runaround*), 1942)

Au niveau du cas concret, la question de la responsabilité morale sera toujours de savoir qui utilise et comment il utilise cet outil et de savoir s'il y a un risque de dommage pour un vivant tiers.

Au niveau de la responsabilité morale collective, la question sera de savoir si cette technologie nouvelle compromet le futur des générations actuelles et à venir et d'adopter à cet égard un principe de prudence[34].

[31] Au 21e siècle il convient de compléter : ou à un autre être vivant.
[32] Même remarque que ci-dessus.
[33] Il faudrait ajouter : blessée directement ou indirectement, p.ex. par la destruction de son milieu naturel.
[34] Dans son ouvrage de référence « Le principe de responsabilité », le philosophe Hans Jonas fait appel à notre responsabilité envers l'Humanité entière sur base du constat selon lequel les progrès technologiques permettent désormais la mise en cause de la vie humaine sur la planète terre. Le principe de responsabilité demande une attitude de prudence envers les hommes de demain (Jonas parle de « Fernliebe » à l'opposition de « Nächstenliebe ») et veillera à préserver leur intégrité.

Les risques humains inhérents au développement de l'*I*A

Il n'y a pas de consensus parmi les experts pour savoir quand et combien l'*intelligence* artificielle sera un vecteur de transformation pour la société tout entière. L'*I*A va-t-elle radicalement changer :

- notre milieu naturel qui, seul, assure notre vie et survie ?
- la vie et le survie des autres êtres avec lesquels nous partageons la planète Terre ?
- notre façon de vivre ensemble en simulant des interventions humaines sur les réseaux sociaux ?
- le monde du travail, en remplaçant une grande partie des travailleurs ?
- notre façon d'être dirigés en influençant les élections et en optimisant la manipulation de l'opinion publique ?
- les modes de la guerre par l'utilisation d'armes « intelligentes » ?
- le monde financier par ses prévisions et conseils dont elle ne supporte jamais les conséquences?
- … ?

À l'image de toutes les mutations technologiques des derniers siècles, il est hautement probable que de telles transformations sont inévitables. Le fait de ne pas savoir comment elles évolueront n'est pas fait pour rassurer. À partir du moment où l'Humanité déléguera de plus en plus de tâches à l'*intelligence* artificielle, celle-ci risque de les accomplir au mieux en ignorant tout ce qui est étranger au but assigné, à commencer par nos aspirations vers le progrès humain, progrès qui reste à redéfinir d'urgence au vu des catastrophes écologiques, sociales et éthiques. C'est à partir de là que se profile le défi spécifiquement éthique que ces machines lancent à l'Humanité pour qui la fin ne doit pas justifier les moyens.

Encadré

À titre d'exemple : le risque fatal d'une arme nucléaire aux mains d'une *I*A

Le principe de la dissuasion nucléaire est basé sur la conviction que les pays en question n'attaqueront pas une autre puissance nucléaire de peur d'être eux-mêmes détruits. Le calcul du risque nucléaire est mesuré à l'aide de la théorie des jeux qui sert à modéliser les probabilités et les risques encourus. Ces calculs sont complexes et une puissance nucléaire pourrait être tentée de laisser à une *intelligence* artificielle le soin de déterminer le moment où il devient justifiable, voire nécessaire, de lancer une attaque nucléaire. Dans cette hypothèse, il est probable aussi que l'adversaire utilise lui-même une *intelligence* artificielle qui, à son tour, aura fait son calcul. En d'autres termes, il est parfaitement imaginable que des

machines *intelligentes* décideraient quand le moment sera venu de détruire la planète et il n'est pas exclu qu'elles se concerteraient à cet effet.

Le remède à ce danger consisterait, de prime abord, à demander à ces machines d'agir en fonction de la plus grande probabilité de préserver à la fois l'Humanité et l'effet de dissuasion[35]. Mais pour cela, il faut être certain que l'ennemi n'aura pas programmé sa machine en fonction de la plus grande probabilité de gagner une guerre nucléaire.

Mais cette imprévisibilité n'est pas le seul problème de fond. Depuis le début de l'ère informatique, l'humain délègue de plus en plus fréquemment ses choix à des machines. Cela pose une ribambelle de questions qui vont de l'endormissement, voire de l'atrophie[36] de nos capacités propres jusqu'à l'inconscience de nos actes[37].

Un des grands problèmes risque de résulter de l'incapacité de l'intelligence de l'humain de s'opposer à l'« intelligence » présumée supérieure de la machine. Reprenons l'exemple cité plus haut du juge qui doit apprécier le risque de récidive d'un délinquant. À supposer qu'il utilise une *intelligence* artificielle pour donner une prévision statistique du risque et à supposer que cette machine informe le juge que le risque de récidive est supérieur à 80%, le juge osera-t-il encore donner une chance au délinquant sur base de sa propre intuition et du cas particulier ? Après tout, chaque juge sait qu'il est sujet à erreur et un juge intelligent est un juge qui doute. Le poids de la recommandation de l'*I*A pèsera beaucoup sur sa décision finale, surtout parce qu'il sait qu'en cas de récidive avérée une décision contraire à la recommandation machinale lui sera reprochée.

Il en ira ainsi avec toutes les décisions prises avec l'aide de l'*I*A : le crédit sera-t-il accordé, le candidat sera-t-il pris, un suspect sera-t-il libéré, un malade soigné, etc. contre l'avis d'une *I*A ? Le contrôle humain pourra-t-il dominer alors que l'erreur toujours possible offre deux options inégales : dans un cas, le décideur infortuné pourra dire qu'il a été induit en erreur par la machine et dans l'autre son erreur sera sans excuse autre que sa faute personnelle.

[35] La théorie des jeux à somme non nulle a fait découvrir, entre autres, au politologue Robert Axelrod, les bénéfices de la coopération : coopérer au lieu de l'emporter serait la stratégie gagnante.

[36] Ne risquons-nous pas de calculer moins bien ou de nous orienter moins facilement dans l'espace que nos grands-parents ou parents parce que ces derniers étaient privés de calculateurs de poche et de systèmes de navigation ?

[37] La conscience morale des humains est très influencée par leur distance aux conséquences de cet acte. Ainsi, l'équipage de l'avion qui a lâché la bombe atomique sur Hiroshima n'avait aucune mauvaise conscience dans le cadre donné. Cela tenait au fait que l'équipage agissait sous une autorité supérieure (cf. expérience de Milgram), mais aussi au fait qu'il pouvait brûler vif quelques 100.000 personnes sans s'en rapprocher physiquement. Le contexte technique (bombe, avion et encadrement militaire) permet l'inconscience de la souffrance infligée à autrui.

Pour mesurer le risque éthique lié à l'*I*A, il ne suffit donc pas de se concentrer sur les insuffisances de la machine, surtout que cette machine n'est pas « tombée du ciel », mais que des humains l'ont développée et mise en marche. Au contraire, les questions soulevées par la délégation de nos décisions quotidiennes vers l'*I*A peuvent être généralisées en soulignant l'insuffisance humaine face à la machine. Car l'*I*A n'est pas intelligente au sens humain ; pour l'instant, elle n'en est qu'un modèle, une imitation imaginée par des créateurs humains. Elle n'est rien qu'un outil. Mais c'est bien nous, les humains, qui risquons de nous en faire une fausse image et de lui donner un pouvoir excessif par paresse, arrogance, désinvolture, *hybris* ?

Encadré

Noam Chomsky et l'*I*A : "plagiarism software"

En mars 2023, Noam Chomsky dit au sujet des chatbots qu'ils sont "'marvels of machine learning', but also the banality of evil, rebooted".[38] En effet, selon Chomsky, "The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question. On the contrary, the human mind is a surprisingly efficient and even elegant system that operates with small amounts of information; it seeks not to infer brute correlations among data points but to create explanations..."
"... Let's stop calling it 'Artificial Intelligence' then and call it for what it is and makes 'plagiarism software' because 'It doesn't create anything, but copies existing works, of existing artists, modifying them enough to escape copyright laws...'." ~ Dr. Noam Chomsky, Dr. lan Roberts, Dr. Jeffrey Watumull, New York Times, March 8, 2023.[39]

Encadré

Les limites fondamentales de l'*intelligence* artificielle à l'exemple de l'Art

L'intelligence proprement humaine est le produit des expériences de la personne, à travers ce que nous avons appelé plus haut les filtres de la connaissance. En font notamment partie toutes les interactions sociales qui sont le propre de l'homme.

Ces dernières années, l'*I*A a pu apprendre à produire des œuvres artistiques. Que ce soit en musique, en littérature ou en peinture, les humains ne sont guère plus capables de distinguer une œuvre humaine de celle d'un robot. Ce pas vers la création est généralement perçu comme une étape fondamentale de l'*I*A vers des talents qui semblaient réservés à tout jamais aux humains.

Mais la question n'est pas de savoir si une *intelligence* artificielle est capable de produire des choses dont un humain peut apprécier la beauté[40]. La question fondamentale est de savoir si une *intelligence* artificielle peut produire des choses qu'elle peut elle-même éprouver comme belles. Rien ne permet actuellement de prêter cette capacité à une machine car le concept-

[38] https://fortune.com/2023/03/08/noam-chomsky-ai-chatgpt-are-not-true-intelligence/
[39] https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html
[40] …étant entendu que la référence à la beauté utilisée ici n'implique pas que la beauté soit une condition nécessaire à l'art.

même de la beauté n'existe pas en dehors d'une conscience capable d'en ressentir la valeur. On peut retrouver dans ces œuvres des modèles ou clés qui ont été variés de façons multiples, un collage ou un assemblage. Généralement ce ne sont pas des créations qui émeuvent, mais elles évoquent plutôt un déjà-vu.

On est tenté, à titre d'analogie, de se référer à l'artiste belge Wim Delvoye qui a construit *Cloaca*, une machine capable de digestion, c'est-à-dire de transformer des vivres – de préférence issus de la haute gastronomie – en excréments. Pourrait-on dire que cette machine peut manger de façon analogue à ce qu'on prêterait à un humain qui physiquement fait une chose semblable ? Certes non. La machine, aux dires-mêmes de l'artiste, est totalement inutile, voire nuisible. *Cloaca* n'est pas un gourmet et ne le sera jamais.

La mise en cause de l'Art ne provient donc pas de ce que la machine serait devenue artiste. Le défi nait du fait que nous humains ne pourrons bientôt plus faire la différence entre l'art original créé par un artiste et l'imitation de l'art issue d'une machine. L'expression culturelle est une forme de communication entre humains sur ce qui leur est essentiel, sur ce qui a une valeur pour eux. En imitant cette communication, l'*intelligence* artificielle pourrait, le cas échéant, lui enlever son originalité, son *aura* (voir ci-dessous) . La confusion des genres porte en elle le risque de détruire l'indispensable authenticité de l'art, une évolution qu'il faut rapprocher du concept du *Kitsch* né avec la reproduction de masse d'imitations pseudo-artistiques.

Ainsi, on entre dans le monde de Walter Benjamin qui dans « Das Kunstwerk im Zeitalter seiner technischen Reproduzierbarkeit »[41] décrit comment l'art reproduit machinalement perd sa dimension spirituelle et avec elle, son *aura*.

On peut avoir une lecture moins pessimiste en considérant l'*intelligence* artificielle dans l'art comme un outil et non pas comme un acteur. Dans cette perspective le répertoire de l'artiste s'enrichit d'une nouvelle méthode et l'IA s'apparente alors à la photographie : c'est l'artiste qui met en scène ce nouveau médium et transpose avec son appui une recherche artistique réelle.

[41] Walter Benjamin: Suhrkamp Verlag; 8. Edition, Frankfurt am Main 2010.

Conclusion

Il convient d'emblée de souligner que la C.N.E. n'a ni la prétention, ni les moyens, dans le présent avis de traiter tous les enjeux éthiques de l'*intelligence* artificielle ou encore de formuler des conclusions et des recommandations inédites. De nombreux avis et chartes relatifs aux défis éthiques de l'*intelligence* artificielle ont vu le jour au cours de la décennie écoulée, émis tant par des comités d'éthique nationaux que par des organismes internationaux et supranationaux. Ne citons, à titre d'exemple, que l'avis fondé et très exhaustif que le *Deutscher Ethikrat* vient de publier en la matière en 2023.

Il y a peu de doutes que l'*intelligence* artificielle, amorce et moteur d'un véritable changement de paradigme, va inévitablement et massivement changer la vie de nous toutes et tous, ainsi que du vivant et de la Terre dans son ensemble, étant entendu que cet impact se ressentira de façon nuancée selon les concerné(e)s. Les applications d'*intelligence* artificielle auront, probablement, une influence considérable sur une multitude de domaines, entre autres de la société au quotidien. Les transformations technologiques toucheront ainsi des domaines aussi variés que l'économie et l'emploi, la médecine et la santé, la justice et la jurisprudence, l'enseignement et la recherche, l'administration, la défense et la (cyber-)sécurité, etc. Dans tous ces secteurs, les défis lancés par l'*I*A peuvent se présenter différemment et dans de nombreux cas, il n'y a pas de problèmes éthiques notables à attendre. La C.N.E. a mis l'accent de ses réflexions sur des questions générales qui sont susceptibles de toucher surtout les applications d'intelligence générales (*Artificial General Intelligence*) et les applications d'intelligence génératives (*Generative Intelligence*) tels les agents conversationnels.

Dans le présent avis, la C.N.E. a d'abord mis en lumière les propriétés spécifiques de l'intelligence humaine, et a, par la suite, abordé les particularités de l'*intelligence* artificielle. Elle a notamment constaté qu'il n'existe, pour le moment, pas de système formel de pensée morale qui permette de développer une *intelligence* artificielle dotée d'une vraie compétence morale. Il n'y pas d'éthique sans contexte. Une section plus approfondie est consacrée à l'agent conversationnel *ChatGPT* (IA générative ou encore *Large Language Model – LLM*), le modèle linguistique qui a catapulté l'*intelligence* artificielle dans la conscience publique.

Pour éviter toute redondance stérile, la C.N.E. s'abstiendra cependant de répéter des recommandations bien connues et suffisamment thématisées dans les avis et prises de positions

susmentionnés. La C.N.E. a, en revanche, opté pour une approche « principaliste », tirant ses conclusions à partir des principes fondamentaux couramment appliqués en bioéthique[42], une façon de procéder qu'elle s'était déjà appropriée dans ses avis précédents.

Les principes d'égalité et de non-discrimination

L'*intelligence* artificielle est biaisée par principe du fait que ses filtres de la connaissance sont rarement neutres. L'*intelligence* artificielle est, en effet, dépendante de l'information dont elle est nourrie et, comme l'information recueillie par un être humain à travers ses cinq sens, l'information dont se sert une *intelligence* artificielle est forcément filtrée, canalisée, préconçue et orientée.

Parce qu'elle est nourrie par notre histoire, par nos idéologies, nos préjugés et nos erreurs, elle a la capacité de les répéter, de les renforcer, voire d'en générer de nouvelles, issus de ses données d'entraînement, ce qui peut conduire à diverses formes de discrimination automatisée, parfois indirectement, envers des groupes de personnes ou d'êtres vivants partageant des caractéristiques similaires.

Ainsi, la C.N.E. souligne la nécessité de prévoir des mesures efficaces pour éviter de reproduire ou d'amplifier les stéréotypes sexistes, racistes, spécistes, classistes et culturels des êtres humains. Elle invite à lutter contre les différentes formes de biais algorithmiques en assurant, par exemple, la mixité, la diversité et la compétence parmi les personnes codant, traitant, et/ou fournissant les données d'entraînement.

Les principes de bienfaisance et de non-malfaisance

La C.N.E. souligne que l'*intelligence* artificielle doit être développée pour le bénéfice de l'Humanité, du bien commun et de la biodiversité. Dans cet ordre d'idées, aussi les formes de vie non humaines doivent être dotées de « droits » intrinsèques les mettant à l'abri d'agissements purement matérialistes et anthropocentriques les réduisant à de simples moyens aux mains des humains. La réponse de *ChatGPT* à la question « Would it be moral to terraform Mars ? » en dit long sur les risques que comporte une « intelligence » artificielle fondée et téléguidée par une Humanité en proie à l'hybris.[43] En adoptant un principe de prudence, les développeurs, fabricants, autorités ou autres parties prenantes doivent minimiser les risques éthiques liés à l'*intelligence* artificielle et mettre en œuvre des mesures adaptées pour la

[42] Cf. les principes de l'éthique biomédicale d'autonomie, de bienfaisance, de non-malfaisance et de justice établis, p.ex., en 1977 par Tom L. Beauchamp et James F. Childress.
[43] https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

prévention de dommages physiques et sociaux. Ainsi, pour garder le contrôle sur l'*intelligence* artificielle et pour limiter et maîtriser les risques liés à son évolution, il est important de mettre en place des mécanismes de sécurité additionnels telle que la supervision humaine, laquelle permet de contrôler les actions et les décisions de l'*intelligence* artificielle par des expert(e)s qualifié(e)s.

Il y a lieu de protéger la vie privée et les données personnelles d'une manière adéquate. La souveraineté des données utilisées par les systèmes de l'*intelligence* artificielle doit, par ailleurs, être respectée.

La C.N.E. met ainsi en garde contre un abus de l'*intelligence* artificielle par des États, des organisations ou des individus privés, et contre le risque de manipulation de l'opinion publique par des *fake news* et/ou *deep fakes* générées par l'*intelligence* artificielle. La multiplication de contrevérités ne trompe pas seulement le public sur les faits inventés, elle ébranle surtout la confiance dans toutes les communications, vraies ou fausses, ce qui représente un danger majeur pour la vie en société, la liberté de presse et le débat démocratique.

L'*intelligence* artificielle ne connaissant pas de frontières géographiques, les régulations respectives devront faire l'objet de coopérations globales.

Le principe d'autonomie / de responsabilité

De plus en plus de prises de décisions quotidiennes seront à l'avenir déléguées à des systèmes d'*intelligence* artificielle. Dès lors, il est primordial de réduire les (potentiels) effets négatifs de cette délégation sur l'autonomie individuelle et/ou la perception de soi. Dans cet ordre d'idées, la C.N.E. met en garde contre une érosion croissante de l'autonomie des êtres humains résultant d'un transfert progressif de tâches vers l'*intelligence* artificielle. Une déresponsabilisation, tout comme une mise sous tutelle, par l'*intelligence* artificielle sont à éviter à tout prix. La C.N.E. estime qu'il est important d'améliorer les capacités humaines plutôt que de les restreindre ou même les remplacer, renforçant de la sorte la dépendance vis-à-vis des technologies de l'*intelligence* artificielle.

Selon la C.N.E., les technologies basées sur l'*intelligence* artificielle ne doivent pas être utilisées pour remplacer les décisions humaines, mais plutôt pour les accompagner et les soutenir, en tant qu'outil d'aide à la décision. Dans ce sens, des décisions finales ne doivent pas être simplement laissées à des algorithmes, surtout dans des secteurs sensibles comme ceux de la santé, de la justice, de l'enseignement, de la recherche, de la défense, etc.

Autonomie et responsabilité sont des concepts se conditionnant de façon réciproque, qui sont étroitement liés et qui vont souvent de pair. La responsabilité peut être caractérisée comme la capacité d'effectuer des choix autonomes. L'incapacité de l'*intelligence* artificielle à assumer ses décisions implique qu'un être humain doit, en dernier lieu, toujours en rester moralement responsable, ce qui implique à son tour une obligation à en protéger les autres humains tout comme les autres êtres vivants.
La C.N.E. souligne que les systèmes algorithmiques ne disposent pas de la *raison*, n'agissent pas d'eux-mêmes et ne peuvent donc pas assumer de responsabilité.

Équité / Justice

Les systèmes d'*intelligence* artificielle devront être conçus de façon à ce qu'ils ne favorisent ni ne défavorisent injustement un individu ou un groupe particulier et qu'ils soient fondés sur des valeurs sociales comme l'équité, la solidarité ou encore la cohésion sociale. Lors de la conception des technologies de l'*intelligence* artificielle, les conditions humaines et sociales devront impérativement être prises en compte et être garanties. Du fait que l'impact de l'*intelligence* artificielle sur le marché de travail risque d'être sévère, la C.N.E. estime que l'organisation du travail devra en tenir compte et faire l'objet d'un dialogue social permanent.

Selon la C.N.E., l'*intelligence* artificielle doit tendre vers la traçabilité, notamment par une documentation de sa programmation et la transparence sur les données d'entrée afin de rendre compréhensibles, dans toute la mesure du possible, les décisions prises par une *intelligence* artificielle qui serait alors plus digne de confiance.

L'ambition éthique tend vers un mieux vivre ensemble, une cohésion sociale, une liberté responsable, une justice pour tous et le progrès humain accessible au plus grand nombre dans le respect des principes du développement durable.

La C.N.E. estime que l'*intelligence* artificielle a le potentiel de renforcer ces ambitions, mais aussi, au contraire, de contribuer à la violation des droits humains, à l'atteinte aux libertés fondamentales et aux institutions démocratiques, et à la destruction de la biodiversité et du milieu naturel. Elle n'est ni bonne, ni mauvaise en soi, mais elle a un potentiel de transformation qui exige de veiller activement à l'utilisation de ses applications dans le bon sens.

L'être humain et la vie sur Terre au centre des préoccupations

Si les membres de la C.N.E. estiment que l'*intelligence* artificielle présente un fort potentiel d'aide et/ou de catalyseur dans certains domaines, elle recommande en même temps de ne pas sous-estimer les risques qui lui sont inhérents. Au final, les êtres humains doivent et devront toujours être en mesure de maîtriser l'*intelligence* artificielle à tout moment afin de minimiser le risque de toute perte de contrôle, le cas échéant en tirant la fiche électrique.

Quels que soient les bienfaits et les bénéfices attendus de la part de cette technologie nouvelle, l'être humain et la vie en général devront toujours rester au centre des préoccupations. Dans cette optique, les membres de la C.N.E. se prononcent en faveur d'un développement de systèmes d'*intelligence* artificielle et de leurs applications centrées sur l'être humain et, plus globalement, sur le *bios*, l'objectif de ces systèmes n'étant pas de remplacer les humains, mais d'améliorer leurs capacités de réfléchir, de prendre des décisions et d'agir par des moyens technologiques. Il s'agit, à l'avenir, de créer un équilibre entre les possibilités techniques de l'*intelligence* artificielle et de ses applications et les intérêts de l'individu entendu comme membre d'une communauté de destin.

La C.N.E. met en garde contre le fait de considérer l'*intelligence* artificielle comme un « sujet » autonome et de s'en faire ainsi une fausse image, l'*intelligence* artificielle n'étant qu'une simulation des performances cognitives de l'être humain calculant en permanence des modèles partir d'un ensemble de données.
Les systèmes d'*intelligence* artificielle n'étant pas dotés de morale ou d'éthique humaine, ils tirent des conclusions exclusivement sur la base de données et d'algorithmes fournis par les programmateurs et les concepteurs dont la traçabilité devrait être fournie et explicitée par ceux-ci dans la mesure du possible. Si la C.N.E. est d'avis qu'une véritable programmation éthique n'est pas possible à défaut de système formel de pensée morale, elle préconise néanmoins que des freins éthiques soient intégrés dans les processus de conception des systèmes algorithmiques à la base de l'*intelligence* artificielle (*ethics by design*).

Cependant, une éthique professionnelle pour les constructeurs et programmateurs de ces systèmes, des instances de contrôle et de réclamation indépendantes – disposant de l'expertise technique, éthique et réglementaire requise contre les abus – ainsi que et des mécanismes évitant une monopolisation des systèmes sur le marché devraient être élaborés au niveau mondial sinon européen.

Dans le même temps, la C.N.E. souligne que les concepts moraux et éthiques peuvent varier selon les contextes et les cultures et que ce fait rend une « programmation éthique » d'office bien complexe. De même, il ne faut pas oublier que l'*intelligence* artificielle est toujours inscrite dans le contexte de pratiques sociales qui peuvent évoluer dans le temps et selon les cultures. C'est la raison pour laquelle la C.N.E. invite à réfléchir à la fois sur l'influence de l'*intelligence* artificielle sur la perception que l'être humain a de soi-même et sur les implications actuelles et futures de l'*intelligence* artificielle sur les sociétés et la vie sur Terre.

Finalement, la C.N.E. se prononce en faveur d'un développement responsable de l'*intelligence* artificielle, de donner le temps à la recherche en matière d'*intelligence* artificielle d'avancer, aux citoyens de se former et à la société de se préparer aux changements à venir.

Encadré

Il suffit de lire l'article ci-dessous de l'Académie française pour comprendre l'*intelligence* artificielle n'est pas (encore ?) intelligente[44] :

Intelligence, nom féminin

Étymologie : xii^{e} siècle. Emprunté du latin intellegentia, intelligentia, « action de discerner, de comprendre ».

I. Ensemble des facultés intellectuelles.

1. Faculté de comprendre, de concevoir, de connaître, et notamment faculté de discerner ou d'établir des rapports entre des faits, des idées ou des formes pour parvenir à la connaissance. L'intelligence humaine. Le développement de l'intelligence. Faites appel à votre intelligence plus qu'à votre mémoire. Exercer son intelligence. Cet homme a l'intelligence vive, lente. Il lui faudra beaucoup d'intelligence pour résoudre ce problème. Absolument. Avoir de l'intelligence, comprendre et raisonner avec aisance, rapidité.

- Par analogie. L'intelligence du cœur, discernement intuitif qui supplée, dans certaines situations, les qualités intellectuelles. Intelligence artificielle, voir Artificiel.
- Par métonymie. Être doué de la faculté de penser, esprit. Ce livre est à la portée de toutes les intelligences. C'est une intelligence supérieure. Être immatériel, purement spirituel. Dieu est la souveraine Intelligence, la suprême Intelligence.

2. Par extension. Aptitude à adapter son comportement à une situation nouvelle, adresse qu'on montre dans une situation donnée, habileté dans le choix des moyens qu'on emploie pour parvenir à un certain résultat. Montrer de l'intelligence, manquer d'intelligence dans la

[44] https://www.dictionnaire-academie.fr/article/A9I1608

conduite d’une affaire. Il s’est acquitté de sa mission avec intelligence. Il a eu l’intelligence de se taire.

■ Se dit aussi pour des animaux. L’intelligence d’un chien, d’un singe.

3. Par métonymie. Connaissance approfondie, compréhension nette et facile qu’on a de quelque chose. Une parfaite intelligence d’un texte. Avoir l’intelligence des affaires, de la politique.

■ Loc. Pour l’intelligence de, pour avoir une bonne compréhension, une connaissance exacte de. Pour l’intelligence de cette affaire, de la chose, il ne faut omettre aucune circonstance dans votre récit.

II. Entente entre deux ou plusieurs personnes.

1. Accord de pensée ou union de sentiments (le plus souvent précédé d’un adjectif). Ils sont en parfaite intelligence. Ce démêlé n’a pas rompu leur intelligence, leur bonne intelligence. Ces voisins vivent en mauvaise intelligence, ne s’entendent pas.

2. Complicité, connivence. Être d’intelligence avec quelqu’un. Ils sont d’intelligence pour vous surprendre, pour vous tromper. Le témoin était d’intelligence avec la partie adverse. Faire à quelqu’un des signes d’intelligence. Être accusé d’intelligence avec l’ennemi, de trahison.

■ Spécialement. Au pluriel. Relations secrètes. Entretenir des intelligences avec une puissance étrangère. Par métonymie. L’assiégeant avait des intelligences dans la place.