Nouvel humanisme à l'ère des neurosciences et de l'intelligence artificielle



# Livre blanc NHNAI : Exploration sociétale de l'humanisme au service de l'éthique et de la gouvernance de l'IA

Première version - janvier 2026

UCLy LYON CATHOLIC UNIVERSITY
UNITÉ DE RECHERCHE confluence SCIENCES ET HUMANITÉS EA1598
IFCU International Federation of Catholic Universities

# Nouvel humanisme à l'ère des neurosciences et de l'intelligence artificielle

## *Livre blanc NHNAI : Exploration sociétale de l'humanisme au service de l'éthique et de la gouvernance de l'IA*

Ce document constitue une première version élaborée par l'équipe de coordination de NHNAI grâce au soutien direct et aux contributions des partenaires de NHNAI. Il est désormais publié et ouvert aux commentaires et réactions. Les retours d'expérience permettront l'élaboration d'une deuxième version mise à jour.

# Préambule

L'humanité est confrontée à des défis majeurs, notamment le changement climatique, les inégalités, les tensions géostratégiques, l'affaiblissement des organisations politiques démocratiques (avec notamment des questions aiguës concernant l'équilibre entre les pouvoirs publics et privés ainsi que l'intelligence collective et la menace de la post-vérité). De plus, ces multiples défis se manifestent tous simultanément et selon une dynamique très rapide et brutale. Comme l'illustre la célèbre analogie avec le pharmakon, l'IA dispose d'un potentiel considérable pour soit contribuer à atténuer ces défis, soit les amplifier.

L'ère actuelle de l'IA — portée par des avancées telles que l'apprentissage profond et les grands modèles linguistiques (LLM) — est marquée par un développement frénétique et une concurrence mondiale intense. Cette ruée est alimentée par la perception de l'IA comme un moteur économique majeur, conduisant à une « tyrannie de la précipitation » où les tentatives de régulation sont dénoncées comme un suicide économique. Cet environnement frénétique, dicté par le marché, associé à des asymétries de pouvoir croissantes favorisant les géants de la tech qui adhèrent souvent à des idéologies techno-solutionnistes, rend le développement éthique de l'IA extrêmement difficile.

Le projet NHNAI repose sur l'idée fondamentale que, pour dépasser cette adhésion aveugle à la technologie et mettre véritablement l'IA au **service de l'humanité**, un effort collectif et intense de renforcement des capacités éthiques doit être mené afin de renforcer le discernement individuel et collectif et de contribuer à un soutien horizontal et ascendant solide à la **gouvernance démocratique de l'IA**. Se contenter de souligner les déséquilibres de pouvoir ne suffit pas. Il faut les combattre. Mais nous pensons qu'orienter de manière adéquate le développement et l'utilisation de l'IA relève de la responsabilité de toutes les personnes concernées. En effet, les asymétries de pouvoir ne sont pas le seul obstacle sur la voie d'une gouvernance démocratique solide de l'IA. Un autre défi consiste à être capable de fixer des objectifs et des finalités à l'IA. Il s'agit là d'une responsabilité collective décisive. Quels sont les projets de société que l'IA devrait servir ? Quelles sont les visions de la nature humaine, du développement et de l'épanouissement qui opéreront en arrière-plan ?

Le présent livre blanc est le fruit des efforts du réseau NHNAI pour contribuer à cet effort collectif nécessaire de discernement. Il propose plusieurs recommandations formulées par les experts universitaires du réseau (notamment sur la base des diverses discussions menées dans les neuf pays ayant participé à la première phase 2022-2025). Les recommandations sont divisées en trois volets principaux : 1) des recommandations pour l'organisation d'une réflexion collective visant à construire un soutien solide à la gouvernance démocratique de l'IA, 2) des recommandations sur les éléments fondamentaux de la compréhension de l'IA ainsi que de ce que signifie être humain (éléments sans lesquels les discussions collectives pourraient être entravées), 3) des recommandations sur des thèmes importants qui devraient être explorés dans le cadre de la réflexion collective.

Avant de proposer une analyse détaillée qui déploie l'intégralité du contenu des recommandations du NHNAI, le document commence par un résumé exécutif présentant les recommandations sous une forme condensée. Le lecteur peut soit consulter ce résumé et se référer occasionnellement à l'analyse détaillée pour plus de précisions, soit commencer directement par l'analyse détaillée.

# Table des matières

## Résumé

**Principales sources d'inspiration ..... 8**

**Une réflexion sociétale collective visant à susciter un large soutien en faveur d'une gouvernance démocratique de l'IA ..... 9**

Recommandation n° 1 : une réflexion inclusive et horizontale ..... 9

Recommandation n° 2 : Une réflexion non seulement sur les utilisations des systèmes d'IA, mais aussi sur leur conception ..... 9

Recommandation 3 : Un effort transformateur de véritable réflexion politique, avec le soutien d'experts ..... 9

Recommandation 4 : Renforcement des capacités éthiques pour des forces motrices ascendantes et horizontales ..... 10

Recommandation 5 : Une réflexion spécifique sur ce que signifie être humain ..... 11

**Étapes fondamentales concernant l'IA et les humains pour soutenir la réflexion sociétale ..... 11**

Recommandation 6 : Garantir une compréhension solide et responsabilisante des technologies d'IA ..... 11

Recommandation 7 : Acquérir des intuitions fondamentales sur les spécificités de l'humain par comparaison avec les machines ..... 14

**Quelques thèmes clés à explorer collectivement ..... 17**

Recommandation 8 : Explorer comment aider et soutenir les humains dans leur rapport à la connaissance et à la vérité ..... 17

Recommandation 9 : Réfléchir à la manière dont l'IA peut servir l'intelligence humaine (collective) ..... 18

Recommandation 10 : Explorer comment l'IA peut contribuer à l'action et à la responsabilité humaines ..... 19

Recommandation 11 : Remettre en question les notions de progrès, de bonne vie et de vulnérabilité ..... 19

Recommandation 12 : Cultiver notre sensibilité à la vie et à l'expérience vécue de manière consciente ..... 20

## Analyse détaillée

**Introduction ..... 23**

**Quels efforts en amont pour soutenir une réglementation démocratique forte ? ..... 25**

Une réflexion inclusive et horizontale (recommandation 1) ..... 25

Une réflexion non seulement sur les utilisations des systèmes d'IA, mais aussi sur leur conception (recommandation 2) ..... 25

Un effort de transformation fondé sur une véritable réflexion politique, avec le soutien d'experts (recommandation 3) ..... 28

Un renforcement des capacités éthiques pour des forces motrices ascendantes et horizontales (recommandation 4) ..... 30

Une réflexion approfondie sur ce que signifie être humain (recommandation 5) ..... 31

**Étapes fondamentales concernant l'IA et les humains ..... 33**

Garantir une compréhension solide et responsabilisante des technologies d'IA (recommandation 6) ..... 34

Acquérir des intuitions de base sur les spécificités de l'humain par comparaison avec les machines (recommandation 7) ..... 40

**Quelques thèmes clés à explorer collectivement ..... 48**

Explorer comment aider et soutenir les humains dans leur rapport à la connaissance et à la vérité (recommandation 8) ..... 50

Réfléchir à la manière dont l'IA peut servir l'intelligence humaine (collective) (recommandation 9) ..... 52

Explorer comment l'IA peut contribuer à l'action et à la responsabilité humaines (recommandation 10) ..... 54

Remettre en question les notions de progrès, de bonne vie et de vulnérabilité (recommandation n° 11) ..... 57

Cultiver notre sensibilité à la vie et à l'expérience vécue en pleine conscience (recommandation 12) ..... 60

**Remarques finales ..... 64**

# Résumé

## Principales sources d'inspiration

Le discernement collectif requis pour une gouvernance éthique de l'IA doit fondamentalement changer d'esprit et d'orientation, en s'éloignant d'une mise en concurrence directe entre les humains et les systèmes d'IA. Nous devons reconnaître que remplacer une personne par l'IA **ne** revient **jamais** à **« ne rien changer ».** Cette décision substitue toujours la richesse de l'expérience vécue d'une personne vivante à un pur mécanisme et à un automatisme. Par conséquent, la meilleure voie à suivre n'est pas de se concentrer uniquement sur la maximisation de l'efficacité de l'IA dans des rôles où les humains pourraient sembler défaillants ou moins performants.

Cela nécessite un effort collectif de discernement, exigeant un engagement politique considérable de la part des communautés humaines. Nous devons changer la question centrale qui guide le développement de l'IA : au lieu de nous concentrer uniquement sur *« Qu'est-ce que l'IA peut faire mieux que les humains ? »*, nous devrions avant tout nous demander : **« Comment la technologie de l'IA peut-elle nous aider à devenir de meilleurs humains ? »** La question ne devrait pas être *« Quelle est notre place en tant qu'êtres humains dans le nouveau monde de l'IA ? »* — comme si nous nous contentions de nous adapter à une force hors de notre contrôle — mais plutôt : **« Quelle est la place de la technologie de l'IA dans notre monde humain, et quelle contribution peut-elle apporter au développement de sociétés plus humaines ? »** Formulée ainsi, il apparaît clairement que le développement de l'IA est fondamentalement une question politique et éthique, et pas seulement technique.

Nous sommes confrontés à un **problème circulaire** crucial : les systèmes d'IA développés et les utilisations adoptées sans discernement peuvent saper les capacités mêmes nécessaires à un jugement éclairé : la pensée critique, le temps d'attention libre, une prise de décision forte et la sensibilité à la vie.

Cela rend l'engagement collectif en faveur des efforts de discernement encore plus crucial. Il est légitime d'être enthousiaste et optimiste face à l'IA. Nous devons l'être. L'IA recèle un potentiel énorme pour nous soutenir dans notre développement et notre épanouissement. Mais ce potentiel ne peut être pleinement réalisé sans l'implication des communautés sociétales, engagées dans des efforts profonds pour cultiver les capacités de discernement, notamment en favorisant une meilleure compréhension de la technologie de l'IA elle-même (culture de l'IA) ainsi que de la nature et de la condition humaines. Il ne fait aucun doute que l'IA peut apporter de merveilleuses contributions à l'épanouissement humain, mais seulement si nous cultivons et gardons toujours une conscience claire du prix que nous sommes appelés à payer en termes d'engagement dans des efforts exigeants de discernement éthique et politique.

## Une réflexion sociétale collective pour construire un soutien solide à la gouvernance démocratique de l'IA

### Recommandation 1 : Une réflexion inclusive et horizontale

Pour garantir **une réglementation** efficace et solide **de l'IA**, la société doit mettre en place un processus **de réflexion inclusif et horizontal**. Si les groupes d'experts et les représentants politiques sont nécessaires, leur confier l'intégralité de l'effort réglementaire risque de créer un cadre technocratique ou élitiste, qui pourrait ne pas aboutir aux meilleurs résultats.

Une véritable gouvernance nécessite la mobilisation **de toutes les parties prenantes concernées**. Une large participation est essentielle non seulement pour la **qualité et** le **contenu** des réglementations, mais aussi pour garantir **une forte sensibilisation sociale** à ces règles. Sans une compréhension et une implication généralisées, **l'application** et le succès à long terme des dispositifs de régulation de l'IA seront gravement compromis.

### Recommandation n° 2 : Une réflexion non seulement sur les utilisations des systèmes d'IA, mais aussi sur leur conception

La gouvernance de l'intelligence artificielle doit fondamentalement aller au-delà de la simple réglementation **des utilisations** pour englober les phases de **conception et** de **développement** des systèmes d'IA. La vision culturelle traditionnelle, héritée de la modernité, sépare souvent **les faits neutres** (technologie/science) **des valeurs** (volonté sociétale/démocratique), ce qui conduit à la croyance simpliste selon laquelle l'éthique ne s'applique que lors du choix de l'utilisation d'une technologie toute faite.

Cette vision est profondément erronée, car une longue tradition de critique technologique montre que les artefacts comportent intrinsèquement **des implications politiques et éthiques**. En particulier avec l'IA, les considérations éthiques commencent bien **en amont** de l'utilisation. Par exemple, les préoccupations concernant les outils de classification biaisés, les systèmes de décision algorithmiques ou les moteurs de recommandation motivés par la capture d'attention démontrent que la « qualité » éthique des résultats est largement **intégrée dans la conception du système**, et n'est pas seulement déterminée par le choix de l'utilisateur.

Par conséquent, la réflexion de fond que la société doit mener doit reconnaître la nécessité d'une délibération éthique **en amont** — dès les **étapes** de **conception et** de **développement** de l'IA. Cela nécessite une contribution participative des citoyens et des parties prenantes, non seulement pour éclairer la législation externe, mais aussi pour intégrer directement des valeurs dans le processus de création grâce à des méthodologies telles que **la conception sensible à la valeur (Value Sensitive Design, VSD)**. Ce n'est qu'en comprenant que l'éthique et la gouvernance doivent imprégner l'ensemble du cycle de vie de l'IA que la société pourra fournir les fondements nécessaires à une réglementation solide et significative.

### Recommandation 3 : Un effort transformateur de véritable réflexion politique, avec le soutien d'experts

Une éthique et une gouvernance efficaces de l'IA doivent s'appuyer sur une **réflexion politique transformatrice** soutenue par des experts, allant bien au-delà de la simple collecte d'opinions publiques. La conception des systèmes d'IA comporte une **dimension politique inéluctable** qui ne peut se réduire à la simple juxtaposition d'opinions existantes que de simples sondages pourraient recenser, ni à l'opinion de la majorité en cas de tendances contradictoires. Cela reviendrait à supposer soit que des réponses et des solutions valables sont disponibles, prêtes à être mises en œuvre, soit (d'une manière plus pessimiste) qu'il n'y a rien de plus à faire qu'une telle collecte. Cela est particulièrement problématique car les préférences sociétales actuelles contribuent souvent aux problèmes mêmes que soulève l'IA (par exemple, des algorithmes biaisés reflétant des préjugés sociétaux, ou l'adoption sans discernement de modèles économiques disruptifs).

Pour aborder les questions complexes soulevées par l'éthique et la gouvernance de l'IA, qui sont souvent **des « problèmes épineux »** mêlant des questions techniques, éthiques et politiques, **des transformations sociales et d' s** sont nécessaires. Cette approche nécessite **une coproduction transdisciplinaire de solutions** où divers experts (philosophes, informaticiens, économistes, etc.) soutiennent la réflexion sociétale collective. Il est essentiel que ce soutien ne soit pas un exercice technocratique. On n'attend pas des experts qu'ils apportent des réponses aux problèmes en jeu. Au contraire, leur expertise doit faciliter les processus **d'autocritique, d'enrichissement mutuel et d'autotransformation** au sein des groupes sociaux, leur permettant ainsi de construire des contributions sociétales solides pour orienter le développement et les utilisations de l'IA.

Cet effort de réflexion collective doit remettre en cause la **dichotomie** trompeuse **entre faits et valeurs**, qui traite à tort les questions éthiques ou politiques comme des questions purement subjectives relevant du libre arbitre. En s'engageant dans **des communautés de discussion et de délibération rationnelles**, les participants à cet effort collectif de réflexion s'acquittent du devoir de rechercher sincèrement la validité et la vérité, allant au-delà d'un simple patchwork d'opinions contradictoires. Cette réflexion rationnelle et collective est essentielle pour établir une compréhension commune des enjeux, en laissant place à des tensions légitimes (comme la vie privée face à la sécurité), et en construisant une base solide pour orienter efficacement le développement et les utilisations de l'IA vers des objectifs éthiques convenus.

### Recommandation 4 : Renforcement des capacités éthiques pour des forces motrices ascendantes et horizontales

Une éthique et une gouvernance solides en matière d'IA nécessitent **le renforcement de capacités éthiques** au sein des communautés sociétales. Ce travail collectif, réflexif et transformateur (recommandation 3) — qui consiste à recenser les enjeux et à rechercher des réponses — permet aux citoyens et aux parties prenantes de contribuer à l'orientation de l'IA.

Cette **autonomisation** est essentielle pour plusieurs raisons :

- Il permet une contribution ascendante à la conception de principes éthiques et de réglementations juridiques **descendants**.

- Il facilite la **contextualisation et l'application** des règles au niveau local, créant ainsi **des forces** d'orientation **ascendantes et horizontales**.
- Il favorise des choix quotidiens et de consommation éclairés, contribuant ainsi à créer **un espace économique viable pour l'entrepreneuriat éthique**.
- Il permet aux communautés de définir **des cas d'utilisation à forte valeur ajoutée**, garantissant ainsi que les technologies d'IA contribuent véritablement aux sociétés humaines.

Un renforcement généralisé des capacités éthiques est essentiel pour orienter efficacement le développement de l'IA.

### Recommandation 5 : Une réflexion approfondie sur ce que signifie être humain

**Une éthique et une gouvernance** efficaces **de l'IA** dépendent de manière cruciale d'une réflexion fondamentale sur **l'humanisme** — plus précisément, sur ce que signifie être humain et ce que nous *voulons* être en tant qu'êtres humains. En effet, de nombreux principes éthiques majeurs de l'IA — tels que le maintien de l'humain **« dans la boucle »** ou la recherche de **l'« épanouissement humain »** — font directement appel à la notion d'humain. Plus fondamentalement, la nature même de l'éthique, définie par des figures telles que Ricoeur comme la recherche d'une **« bonne vie » avec et pour les autres au sein d'institutions justes,** exige de réfléchir à ce que signifie être humain pour forger le sens de ces termes.

L'essor de technologies telles que l'IA, qui confère aux humains le pouvoir d'**une auto-modification en profondeur** (anthropotechnè), rend plus nécessaire que jamais une boussole fiable concernant notre nature humaine.

Cependant, la notion d'humanisme n'est ni claire ni consensuelle, faisant face aux critiques des courants anti-, post- et transhumanistes qui remettent notamment en question son accent problématique sur l'autonomie ou son rôle dans le mythe du « maître et possesseur de la nature ». Au lieu d'un rejet pur et simple, la recommandation est de s'engager dans une **exploration renouvelée** de l'humanisme. En renouant de manière critique avec des idées fructueuses — comme l'accent central mis par Kant sur **la liberté, la responsabilité et la faculté de jugement** — et en les associant aux connaissances issues de l'IA et des sciences cognitives, nous pouvons favoriser une **exploration collective** visant à esquisser les contours d'un **nouvel humanisme**. Cette compréhension approfondie et partagée de notre propre nature humaine (de qui nous sommes ainsi que de qui nous devrions être) est indispensable pour élaborer des lignes directrices éthiques fiables pour l'IA.

## Étapes fondamentales concernant l'IA et les humains pour soutenir la réflexion sociétale

### Recommandation 6 : Garantir une compréhension solide et responsabilisante des technologies d'IA

Actuellement, l'IA est souvent présentée à tort comme une vague illimitée et inexorable menant vers l'AGI ou la SAI, alimentant des discours sur l'obsolescence humaine ou la perte de contrôle. Ces représentations trompeuses ignorent la dimension cruciale de l'IA en tant **qu'artefact humain**, occultant sa réalité politique et éthique. Il est donc essentiel de garantir une compréhension pas trop abstraite des technologies d'IA au sein des communautés participatives. Cette culture est la condition préalable nécessaire pour que les citoyens et les parties prenantes puissent efficacement construire, débattre et orienter les projets sociétaux positifs que l'IA devrait servir.

**Démystifier l'apprentissage automatique**

L'apprentissage automatique (ML) est une recherche technique sophistiquée visant à trouver le **paramétrage** optimal **d'une architecture informatique** pour exécuter des tâches qui résistent à la programmation conventionnelle, étape par étape (comme la classification d'images complexes). Les ingénieurs conçoivent une architecture spécifique comportant de nombreux types d'opérations et d'agencements de celles-ci, spécifiés par des paramètres libres (par exemple, des coefficients). Ils écrivent ensuite un programme par essais et erreurs (ou similaire) pour ajuster ces paramètres, guidés par **un retour d'information** explicitement défini.

Ce niveau minimal d'explication est **essentiel pour démystifier l'IA et gérer les attentes**, car il montre qu'il **n'**existe **pas d'IA unique, gigantesque et toute-puissante**. Il existe au contraire **diverses techniques** qui dépendent fortement de l'architecture (comme les réseaux convolutifs ou les transformateurs) et, surtout, de la nature du retour d'information fourni :

1. **Nouveauté et nouvelles façons de résoudre :** dans des configurations étroites et spécialisées, comme les jeux de société, les jeux vidéo ou d'autres environnements simulés, le retour d'information peut être défini mathématiquement. Cela permet au système de trouver éventuellement **des solutions** véritablement **novatrices** et d' r de meilleures façons d'atteindre l'objectif représenté par le retour d'information. Ici, les systèmes *peuvent* produire des résultats dont les humains n'avaient pas conscience, mais il s'agit de contextes de niche.

2. **Reproduction et généralisation limitée :** dans les applications les plus répandues (comme les LLM et la classification d'images), le retour d'information est défini par comparaison avec un vaste ensemble de **données d'exemples**. Dans ces cas, l'objectif principal du système est d'imiter et de corréler, et non d'innover. Il est **trompeur d'espérer des résultats radicalement nouveaux** ; le système est récompensé pour reproduire le passé, et non pour inventer l'avenir. Par conséquent, la capacité de ces systèmes à généraliser à de nouveaux problèmes est très limitée.

De plus, cette vision un peu moins abstraite permet de mieux anticiper la fiabilité des différents systèmes. Elle permet d'atténuer la tendance répandue à décrire les LLM et autres programmes d'apprentissage automatique (profond) comme des boîtes noires impénétrables. En fait, les opérations effectuées par ces programmes sont connues. Il est en principe possible d'examiner les valeurs des différents paramètres libres et les calculs auxquels ils aboutissent. Le problème est que cela n'apporte que peu d'informations sur les raisons pour lesquelles un programme

donné fonctionne bien ou non. Nous manquons souvent d'une **théorie** fiable **de l'erreur**, ce qui entraîne la possibilité d'« hallucinations » imprévisibles. Mais, en matière de fiabilité encore une fois, on peut distinguer différents types de systèmes. Par exemple, la fiabilité peut être raisonnablement évaluée et donc attendue dans des outils spécialisés et testables empiriquement (par exemple un programme spécialisé dans la classification d'images médicales), mais nos attentes devraient diminuer considérablement dans le cas de l'IA générative à usage général, dont les utilisations devraient être adaptées en conséquence.

**Rappel de la matérialité de l'IA**

L'IA et la technologie numérique sont souvent présentées à tort comme immatérielles. Cependant, tous les programmes s'exécutent sur **du matériel** physique conçu pour **la transformation automatique et mécanique** de configurations matérielles (par exemple, des orientations magnétiques, des états électroniques) auxquelles les humains ont attribué une signification (comme des 0 et des 1, des mots ou des chiffres). À cet égard, l'ordinateur peut être considéré comme l'aboutissement d'une longue histoire des technologies de l'information, remontant à l'invention même de **l'écriture** (qui consiste précisément à donner des significations à des formes matérielles particulières et constitue la toute première étape pour construire par la suite des automates qui agiront sur ces formes).

Une propriété fondamentale de l'ordinateur réside dans son **inertie intrinsèque et imperturbable**. Il traite l'information en manipulant de manière efficace et précise ces configurations matérielles selon un programme. Les ordinateurs ne renferment pas en soi de sens, d'émotions ou de conscience ; ce sont simplement **des machines extraordinaires** qui agissent sur des configurations qui ont un sens pour nous afin de créer mécaniquement de nouvelles configurations, que *nous* interprétons ensuite comme du texte et des images (pouvant éventuellement exprimer des sentiments).

**Souligner la dépendance directe à l'égard de l'intelligence humaine**

Il est essentiel de **s'émerveiller des succès de l'IA pour les bonnes raisons** : ils démontrent la capacité de l'humanité à construire des mécanismes inertes et complexes qui *simulent* un comportement intelligent. L'IA est fondamentalement un produit de l'intelligence humaine et en **dépend** de manière **irréductible**.

La notion d'une Super Intelligence Artificielle (SAI) magique produisant des résultats réels que nous ne pouvons vérifier est illusoire. Les humains restent entièrement responsables de la construction des systèmes et de l'évaluation de leurs résultats. Des niveaux élevés d'intelligence humaine sont requis tout au long du processus :

- **Conception :** elle nécessite des programmeurs et des ingénieurs compétents pour créer des architectures informatiques et des procédures d'apprentissage.
- **Accompagnement :** il faut des experts du domaine pour définir formellement les mécanismes de rétroaction, et/ou de nombreux humains intelligents pour fournir de bons exemples pour les ensembles de données d'entraînement (par exemple, des journalistes, des chercheurs, toute personne partageant du contenu sur des systèmes d'information).

- **Évaluation :** L'effort et l'intelligence humains sont nécessaires pour instaurer la confiance, évaluer la qualité et définir les limites de fiabilité des systèmes. Il faut reconnaître l'existence d'un véritable problème de « contrôle » lorsque la fiabilité des systèmes n'est pas suffisamment garantie (en particulier avec les puissants systèmes d'IA générative dont les résultats pourraient devenir très difficiles à anticiper et à sécuriser).

Le **plus haut niveau d'intelligence humaine collective** est requis pour la tâche cruciale qui consiste à définir **des orientations et des objectifs** adéquats pour l'IA. Il faudra également beaucoup de force et d'intelligence humaines pour s'abstenir d'utiliser, dans des contextes insuffisamment sécurisés, des systèmes qui n'offriraient pas de garanties suffisantes de fiabilité.

Recommandation 7 : Sécuriser certaines intuitions fondamentales sur les spécificités des humains par comparaison avec les machines

Une éthique et une gouvernance efficaces de l'IA nécessitent une réflexion sur les spécificités humaines par rapport aux machines. Les affirmations disruptives ou prophétiques suggérant que les machines pourraient (bientôt) posséder des traits humains fondamentaux tels que la conscience ou le libre arbitre doivent être considérées avec une extrême prudence. S'il faut résister à un dogmatisme aveugle, il est loin d'être évident qu'il existe actuellement de bonnes raisons de réviser en profondeur nos intuitions fondamentales sur les spécificités des humains par rapport aux machines.

**Résister à l'injonction de l'« outcomisme »**

Les communautés engagées dans la réflexion sur l'éthique et la gouvernance de l'IA doivent **résister** activement **à l'« outcomisme »,** cette prescription qui consiste à restreindre la discussion sur les traits humains de l'IA (conscience, intentionnalité, libre arbitre) à la seule comparaison **des résultats** observables entre les systèmes d'IA et les humains.

Cette mentalité restrictive trouve son origine dans une remise en cause épistémique de l'introspection, s'inscrivant dans la lignée du fonctionnalisme et du behaviorisme, qui considèrent la vie intérieure comme un « problème difficile » inaccessible à une étude scientifique objective. L'outcomisme peut par exemple conduire à nier la possibilité même d'une distinction entre une expérience vécue authentique de la compassion et la simple manifestation d'un comportement compatissant.

Le danger réside dans le fait que, alors que l'IA générative excelle à passer **des tests de type Turing** (par exemple, en créant des œuvres d'art impossibles à distinguer de celles d'un humain ou en imitant des experts en éthique), cette focalisation sur les résultats ébranle rapidement des intuitions humaines bien ancrées, risquant ainsi de remettre en cause la légitimité de prendre en compte l'expérience humaine vécue en amont des résultats auxquels nous sommes confrontés (des facteurs tels que la présence d'un artiste humain qui a élaboré une image que nous contemplons).

Pour conserver un discernement éthique solide, nous devons éviter cette mainmise sur le pouvoir. S'il est salutaire de remettre en question les croyances fondamentales, réduire le débat

complexe sur la conscience et l'agentivité (humaines) à une simple comparaison des résultats ne mobilise qu'une fraction limitée des arguments (philosophiques) disponibles.

**Prendre au sérieux l'expérience vécue, la vie et la biologie**

Pour éviter de brouiller de manière trompeuse la distinction fondamentale entre les humains et les machines, nous devons remettre en question l'hypothèse centrale selon laquelle les expériences intérieures vécues sont des sortes de « trous noirs » sur lesquels nous ne pouvons rien dire d'objectif ou de fiable (comme le fait l'« outcome-ism »).

Nous pouvons et devons résister au rejet systématique de l'introspection. L'expérience de pensée de la « chambre chinoise » de John Searle, par exemple, légitime le recours à **l'expérience introspective** pour réfuter les affirmations universelles sur la nature computationnelle de l'esprit.

De plus, nous devons remettre en question l'affirmation selon laquelle la conscience phénoménale est uniquement un « problème difficile » scientifiquement (objectivement, sérieusement) abordable uniquement par le biais de modèles computationnels visant à reproduire les connexions entre les entrées et les sorties. Des approches comme celle d'Antonio Damasio démontrent la possibilité d'élargir le champ d'étude, en étudiant la conscience non seulement à travers les propriétés computationnelles du cerveau, mais aussi à travers son ancrage dans les **mécanismes biologiques sous-computationnels** et l'interaction riche et organique entre le système nerveux et le reste du corps vivant (notamment via **l'interoception**, notre perception interne de notre propre corps).

En prenant **au sérieux la vie et la biologie**, nous pouvons distinguer les êtres vivants des simples artefacts de la technologie de l'information. Cette perspective renforce nos intuitions fondamentales concernant ce qui est vivant, conscient et possède une autonomie ou un libre arbitre. Ces intuitions sont loin d'être des préjugés démodés et obsolètes que des approches scientifiques sérieuses pourraient déjouer de manière convaincante. Au contraire, lorsqu'elles ne sont pas réduites de manière illégitime à l'outcomisme, ces investigations scientifiques confirment plutôt que nos intuitions peuvent servir de base fiable pour une exploration plus approfondie.

**Assurer une compréhension solide des spécificités fondamentales de l'être humain**

Nous devons aborder de manière critique la tendance linguistique courante à attribuer des traits humains tels que **« l'intelligence », la « prise de décision »** et un **« rapport à la vérité »** à l'IA et aux systèmes numériques. Si l'utilisation de ces termes pour décrire des fonctions automatisées peut être pratique et admissible, elle risque d'éroder le **statut ontologique** unique des humains, nous encourageant à nous considérer comme de simples machines. Une gouvernance éthique exige que nous **préservions la possibilité** que ces termes désignent une réalité plus profonde et irréductible, enracinée dans la vie et l'expérience humaine vécue.

## Autonomie et prise de décision

La prise de décision et l'autonomie des machines, même dans le cas d'une IA avancée, ne sont valables que dans un sens restreint. Les ordinateurs sont fondamentalement **mécaniques et**

**inertes**, fonctionnant comme des systèmes déterministes où la complexité de la réponse est le seul élément qui les différencie d'un thermostat. Ils opèrent en adhérant rigoureusement à des algorithmes (potentiellement incroyablement complexes) et à des données passées, réagissant toujours de la même manière à une même entrée dans un état donné.

À l'opposé, **l'autonomie humaine** est un concept ontologiquement plus solide, profondément ancré dans **la vie et les phénomènes biologiques**. La prise de décision humaine, que nous connaissons intimement grâce à l'introspection , transcende toute procédure mécanique. Elle se définit par la **capacité à s'écarter** des régularités passées et à réagir différemment à des sollicitations identiques. Cette capacité au choix volontaire et à **l'autonomie pratique** est indissociable de notre constitution biologique, ainsi que de notre expérience phénoménale vécue, qui englobe notre **vie affective et émotionnelle**.

Cela revêt une importance capitale dans le domaine moral. La prise de décision morale humaine n'est pas une simple réapplication de réponses passées ; c'est un **pouvoir de créer de la nouveauté.** Cette capacité créative à prendre de la distance par rapport aux normes établies et à envisager véritablement de nouvelles options est indispensable pour reconnaître la possibilité essentielle pour une personne de changer et de laisser son passé derrière elle (une composante clé liée à la dignité humaine). Cette ouverture et ce pouvoir de s'écarter sont au cœur de la prise de décision au sens fort et humain du terme.

### Relation à la connaissance et à la vérité

Il semble indéniable que les systèmes d'IA peuvent produire des résultats vrais (les systèmes les plus puissants sont capables de le faire dans des domaines et des sujets de plus en plus variés). Cela suffit-il pour leur attribuer un rapport à la connaissance et à la vérité ?

Dans la conception philosophique traditionnelle, la connaissance est considérée comme **une « croyance vraie justifiée »,** ce qui exige de bonnes raisons et une justification allant au-delà de la simple production d'énoncés vrais. Sur cette base, certains pourraient affirmer que les machines ont une longueur d'avance sur les humains, car elles n'appliquent que des opérations logico-mathématiques à des données brutes. Elles seraient ainsi dotées d'une sorte d'objectivité parfaite, une forme supérieure de rationalité libérée par principe de toute arbitraire ou subjectivité.

Il est crucial de mettre en garde contre une telle vision déformée, bien que répandue, de la rationalité ou de l'intelligence comme une sorte d'« objectivité mécanique » — un processus purement algorithmique libéré de toute subjectivité. En réalité, l'histoire et la philosophie des sciences révèlent les limites de telles approches : le processus de production de la connaissance, même scientifique, implique un **espace irréductible de liberté** et l'activité inéliminable du **jugement informel** par le sujet connaissant. Il n'existe pas de données brutes neutres ; les jugements et arbitrages humains sont indispensables aux choix méthodologiques et aux intuitions fondamentales. Par conséquent, l'intelligence humaine implique non seulement d'appliquer des critères, mais aussi **de juger de la qualité de ces critères**.

Entretenir une relation avec la connaissance et la vérité au sens humain fort implique une activité critique et réflexive, qui est fondamentalement enracinée dans **l'expérience** humaine

**vécue**. Elle est intimement liée à **l'autonomie** et à **la prise de décision**, car elle requiert la capacité de **prendre du recul** et d'imaginer que les représentations et croyances admises pourraient être différentes. Seule cette capacité à prendre du recul rend les humains sensibles à l'appel à faire un usage responsable de leur liberté et de leur autonomie pratique dans une quête sincère de la vérité.

## Quelques thèmes clés à explorer collectivement

L'éthique et la gouvernance de l'IA sont confrontées à un défi qui va bien au-delà de la simple atténuation des asymétries de pouvoir entre les nations, entre les secteurs public et privé, ou entre les géants de la technologie et les utilisateurs. Il est dangereux de supposer que l'orientation de l'IA est évidente, comme si le seul enjeu consistait à neutraliser les acteurs malveillants.

Une tâche considérable et déterminante de la réglementation de l'IA consiste en un effort collectif visant à définir et à articuler les **buts et objectifs** que l'IA devrait servir, en rejetant le discours selon lequel l'IA (ou la super-IA) est un but incontestable en soi. L'IA peut être utile, notamment en contribuant à résoudre des problèmes civilisationnels complexes, mais seulement si nous **définissons** d'abord **ce que nous attendons d'elle** et affinons nos objectifs humains.

Par conséquent, notre attention doit passer de **la question : « Quelle est notre place en tant qu'êtres humains dans le nouveau monde façonné par l'IA ? »** à la question essentielle : **« Quelle est la place de l'IA dans le monde humain que nous voulons construire ? »** Cela nécessite d'explorer de nombreux sujets différents. Parmi ceux-ci, nous souhaitons mettre en avant et encourager l'exploration collective du défi particulièrement crucial consistant à déterminer comment positionner l'IA pour qu'elle préserve, voire favorise, l'épanouissement des spécificités fondamentales de l'être humain.

> Recommandation 8 : Explorer comment aider et soutenir les humains dans leur rapport à la connaissance et à la vérité

Les humains possèdent la caractéristique fondamentale, bien que faillible, d'entrer en relation avec la vérité et de construire collectivement la connaissance — définie comme **une croyance vraie et justifiée**. Des efforts de discernement approfondis sont nécessaires pour déterminer comment la technologie de l'IA peut véritablement **soutenir** ce processus sans le compromettre.

Pour de nombreux composants des technologies de l'information (tels que les encyclopédies ou revues en ligne, les traitements de texte ou les tableurs), nous nous appuyons sur **de** solides **réseaux collaboratifs** où nous déléguons une grande partie de l'effort d'évaluation de la qualité à des experts humains dignes de confiance (tels que des journalistes, des rédacteurs d'encyclopédies ou des développeurs de logiciels) qui garantissent la fiabilité de la technologie et des résultats qu'elle nous présente. Cette coopération conduit, grâce à la division du travail, à la délégation et à la confiance entre les humains, à un merveilleux environnement numérique

qui rassemble et rend accessibles des corpus de connaissances étonnants (des croyances que nous avons, directement ou indirectement, de bonnes raisons de considérer comme vraies).

**L'IA générative marque une rupture décisive à cet égard.** L'IA ne pouvant se justifier d'elle-même, seule l'autonomie humaine peut en fin de compte juger si les raisons sont suffisantes. Et les LLM fournissent des résultats qui **ne** sont **pas garantis par l'expérience cognitive d'un être humain**. Aucun sous-groupe spécifique ne vérifie le contenu particulier fourni, ce qui fait peser toute la charge de l'évaluation sur l'utilisateur final.

Par conséquent, les communautés humaines doivent acquérir **une culture de l'IA** en ce qui concerne ces questions liées à la connaissance et à la vérité. Cette culture est nécessaire pour distinguer les systèmes existants capables de fournir **des connaissances fiables** (validées par des experts humains) de ceux qui ne sont que de puissants outils d'exploration. Elle est également indispensable pour explorer le type de nouveaux systèmes que nous pourrions développer afin d'apporter un soutien supplémentaire et spécifique aux différentes facettes de nos vies cognitives et épistémiques. Une réflexion approfondie s'impose pour attribuer à l'IA la place qui lui revient afin de préserver, de favoriser et de prolonger nos efforts collectifs pour nous rapprocher de la vérité.

| Recommandation 9 : Réfléchir à la manière dont l'IA peut servir l'intelligence humaine (collective) |
|---|

Pour garantir que l'IA serve véritablement l'humanité, nous devons aller au-delà de la simple fiabilité des outils et nous pencher sur la préservation et le développement de **l'intelligence humaine**. Cela implique une approche à deux volets concernant les compétences individuelles et notre environnement collectif.

Premièrement, nous devons réfléchir aux risques de **déqualification**. Si déléguer des tâches à l'IA peut être efficace, une utilisation excessive peut entraver le développement cognitif. Une analyse approfondie est nécessaire pour identifier quelles compétences et expériences vécues sont indispensables — non seulement pour vérifier les résultats de l'IA, mais aussi pour maintenir les facultés cognitives plus larges nécessaires à l'épanouissement humain.

De plus, l'IA est devenue une **« extension cognitive »** de l'esprit humain. Les algorithmes prédictifs et l'IA générative contribuent à l'éditorialisation de notre réalité, façonnant le « substrat informationnel » sur lequel nous pensons. Parce que la connaissance humaine est fondamentalement **collective** – reposant sur une « décence commune » (la volonté de juger et de savoir en commun, en s'engageant à la validité des croyances devant les autres) –, la manière dont l'IA structure cet environnement est cruciale.

Actuellement, la contribution de l'IA à cette vie intellectuelle collective est préoccupante. Poussés par un modèle économique fondé sur **la captation de l'attention**, les systèmes d'IA génèrent souvent **des bulles cognitives** et **des chambres d'écho**. Cela provoque un « préjudice épistémique », dégradant la confiance et la bienveillance nécessaires à une vie intellectuelle collective saine.

Cette toxicité n'est pas une fatalité. L'IA recèle un immense potentiel pour élargir les perspectives et favoriser la compréhension mutuelle et la décence commune. Cependant, la réalisation de ce potentiel exige un **effort politique et collectif** profond. Nous devons aller au-delà de la simple « correction » ou de la réglementation des algorithmes actuels depuis l'extérieur et **réorienter** fondamentalement **les modèles économiques** pour les éloigner de la capture d'attention, en concevant l'IA de manière à ce qu'elle serve de terreau fertile à une véritable intelligence collective humaine.

Recommandation 10 : Explorer comment l'IA peut contribuer à l'action et à la responsabilité humaines

L'essor des systèmes d'IA complexes, en particulier de **l'IA agentique** capable d'exécuter des actions, nécessite de se concentrer de manière proactive sur la garantie **d'un contrôle humain significatif** — un concept plus large que la simple résolution **des lacunes en matière de responsabilité** juridique. Le véritable problème réside dans la prévention d'un **vide de contrôle** où des systèmes mécaniques puissants et imprévisibles se désalignent des valeurs et des objectifs humains. Cela nécessite **une** solide **culture de l'IA**, permettant aux communautés de discerner quels outils sont fiables pour l'automatisation et lesquels (comme l'IA générative) exigent une supervision stricte pour prévenir des résultats dangereux et imprévisibles.

Une menace éthique profonde pèse sur cette agence : la **tentation de l'externalisation cognitive**. Poussés par la « honte prométhéenne » (un sentiment d'infériorité face aux machines) et par une aversion pour le risque, les humains pourraient déléguer leurs décisions à l'IA afin d'échapper au poids de la responsabilité et à la possibilité d'erreur. Cet abandon est incompatible avec **la prise de décision au sens strict du terme**, qui repose sur la capacité à **s'affranchir** des régularités du passé, à faire preuve de créativité et à imaginer d'autres possibilités d' . Peut-être plus menaçant qu'une prise de pouvoir brutale, nous sommes confrontés au risque d'une **« perte progressive de pouvoir »** — une érosion graduelle de l'influence humaine sur les systèmes sociétaux.

Pour lutter contre cela, un effort collectif est nécessaire afin de résister à la tentation de se réduire à un objet inerte, illusoirement déchargé de toute responsabilité ou de l'exercice pénible de l'autonomie. Nous devons concevoir des systèmes qui **renforcent** le jugement humain **plutôt que de** le **remplacer**. Nous devons débattre de questions clés : **quel prix sommes-nous prêts à payer pour défendre l'autonomie humaine ?** Comment garantir que les algorithmes soutiennent et renforcent nos solides capacités de prise de décision, plutôt que d'encourager un transfert illégitime de responsabilité ?

Recommandation 11 : Remettre en question les notions de progrès, de bonne vie et de vulnérabilité

Pour garantir que l'IA serve véritablement la « bonne vie », nous devons résister **aux raccourcis techno-solutionnistes** qui associent sans discernement l'innovation technologique au progrès humain authentique, en partant du principe que chaque problème a une solution technologique. Cette mentalité risque de réduire les enjeux humains à des indicateurs d'efficacité, en ignorant les causes profondes (telles que les sources sociétales de la solitude

ou de l'épuisement professionnel) au profit de solutions technologiques superficielles. De tels raccourcis, ancrés dans une sorte de **« paradigme technocratique »,** réduisent la réalité à des indicateurs manipulables, considérant le progrès humain uniquement en termes **d'efficacité** et de performances mesurables.

Cette mentalité mécaniste est problématique car elle mutile la quête légitime de liberté humaine en traitant toute limite, erreur ou **vulnérabilité** comme un défaut à éliminer. Cette quête d'infaillibilité et de pouvoir illimité est viciée sur le plan épistémologique et moral.

Comme établi précédemment, **la connaissance** humaine **et la prise de décision** sont intrinsèquement faillibles, nécessitant une **marge de manœuvre** essentielle et la possibilité de commettre des erreurs. La liberté de choisir s'accompagne du risque de faire le mauvais choix. « Améliorer » les humains en supprimant cette faillibilité par l'automatisation ne renforce pas l'action humaine ni la relation à la vérité ; cela les éradique. Le véritable progrès réside dans l'affinement de notre responsabilité et de notre esprit critique, et non dans une quête illusoire d'infaillibilité.

Plus fondamentalement, il est crucial de reconnaître la subtilité de la notion de vulnérabilité. La vulnérabilité ne peut être réduite uniquement à ses aspects négatifs (« vulnération » comme une blessure ou une maladie). La vulnérabilité correspond également à la **possibilité** fondamentale **d'être affecté**. Si nous avons le devoir de prévenir les blessures, tenter d'éliminer la vulnérabilité est une erreur. La robustesse absolue est la négation de la vie ; c'est notre vulnérabilité qui nous permet d'aimer, de ressentir de la joie et de créer des liens avec les autres.

Par conséquent, nous devrions nous détourner de la question trompeuse de savoir si l'IA peut nous rendre tout-puissants, infaillibles et invulnérables. Au contraire, la réflexion collective doit se poser la question suivante : **comment pouvons-nous développer et utiliser les systèmes d'IA pour nous aider à mieux apprivoiser et équilibrer la vulnérabilité et la faillibilité ambivalentes mais essentielles qui se trouvent au plus profond de notre être ?**

| Recommandation 12 : Cultiver notre sensibilité à la vie et à l'expérience vécue consciemment |
|---|

Pour orienter correctement le développement et l'utilisation de l'IA, nous devons **cultiver notre sensibilité à la vie et à l'expérience vécue en pleine conscience**, en évaluant activement quand la **présence d'une personne authentique et vulnérable** est indispensable.

Un tel discernement nécessite un effort collectif considérable et peut s'avérer extrêmement difficile. Il exige notamment de résister à la philosophie de **l'« outcomisme »,** qui menace cet effort en se concentrant exclusivement sur les résultats et les comportements extérieurs. Le « outcomism » peut conduire à une justification problématique de l'utilisation de l'IA pour des rôles sociaux en assimilant **une apparence** convaincante à **une présence** authentique (les communications médicales générées par des modèles de langage grand public pourraient suffire car elles possèdent une empathie apparente, l'art généré par l'IA est légitime s'il peut se faire passer pour de l'art créé par des humains, les compagnons IA pourraient aider à faire

face à la solitude encore mieux que les humains car ils n'abandonneraient ni ne s'opposeraient jamais à leurs utilisateurs…).

Il est indispensable et parfaitement légitime de résister à de telles approches réductionnistes. Comme déjà exposé, une présence humaine authentique est indispensable à l'élaboration des connaissances et à la prise de décision, avec le besoin irréductible d'une utilisation responsable de la liberté. Cela semble primordial dans des contextes tels que les soins de santé et la psychothérapie, où l'affectivité partagée est centrale pour la connexion et l'efficacité des résultats thérapeutiques. On pourrait citer des exemples encore plus évidents : le dessin imparfait d'un enfant, fruit d'heures d'efforts et d'intention, a plus de valeur qu'une image parfaite générée instantanément par une IA. Les paroles de compassion adressées à une personne mourante n'ont aucune valeur si elles ne traduisent pas une expérience authentique et vécue de la compassion.

Ainsi, lorsqu'on s'efforce de déterminer où déployer des systèmes d'IA et à quelles fins, il est parfaitement légitime (et absolument indispensable) non seulement de tenir compte de la qualité « objective » des résultats, mais aussi de réfléchir en profondeur à la valeur potentielle de la **présence d'une personne authentique, vulnérable et sensible** dans les processus d'élaboration en amont. Cela implique que les personnes doivent être informées lorsqu'elles interagissent avec des systèmes qui imitent de manière convaincante les humains (transparence et divulgation de l'utilisation de l'IA), car cette connaissance est essentielle pour permettre une réflexion sur la valeur de la présence authentique.

Veiller à ce que l'IA soit véritablement au service des relations humaines et de la sensibilité à la vie imposera donc de se confronter à des cas ambigus et subtils. Le cas des compagnons IA illustre bien l'ampleur des difficultés. Il semble clair qu'ils ne peuvent remplacer les relations authentiques. Ces dernières nécessitent la présence d'une personne humaine authentique dotée de son autonomie, ce qui donne toute sa valeur à l'interaction, malgré (ou plutôt en raison de) le risque de friction et d'abandon qu'implique une telle liberté. On pourrait toutefois faire valoir que, utilisés comme de simples jouets (à l'instar des personnages imaginaires des livres, des films ou des jeux vidéo auxquels on peut s'attacher), les compagnons IA deviennent inoffensifs. Mais la situation est peut-être plus subtile en raison du niveau d'imitation humaine que ces systèmes peuvent atteindre, ce qui fait courir le risque **d'inculquer aux utilisateurs la négation de l'autre** et de favoriser une culture qui considère l'intimité comme une marchandise schématisée, rendant potentiellement les utilisateurs moins tolérants à l'égard de la véritable autonomie et des frictions inhérentes aux relations humaines authentiques. Dans le même ordre d'idées, proposer de déployer des systèmes d'IA pour soulager des professionnels surchargés (dans le contexte des soins de santé par exemple) pourrait être un moyen de réduire l'épuisement des professionnels, mais cela pourrait aussi devenir un moyen d'éviter de s'attaquer aux causes profondes des difficultés rencontrées.

En résumé, la réflexion collective visant à orienter le développement et l'utilisation de l'IA doit se pencher sur ces questions cruciales : comment, à l'ère de l'IA et, si possible, avec l'aide de celle-ci, favoriser et cultiver notre sensibilité à la vie et à la présence de personnes réelles dans leur expérience vécue ? Plus encore, nous devons cultiver notre **capacité à évaluer la valeur déterminante de la présence authentique par rapport à la simple apparence**, à discerner

où et comment celle-ci est primordiale, et comment elle doit être mise en balance avec des considérations relatives à la qualité des résultats.

# Analyse détaillée

## Introduction

En matière d'IA, dire que ces dernières années ont été frénétiques serait un euphémisme. D'un point de vue technique, l'essor de l'apprentissage profond, récemment renforcé par la technologie des transformateurs, a conduit à de nombreuses réalisations remarquables. Parmi celles-ci, les modèles linguistiques à grande échelle (LLM) et l'IA générative se sont répandus dans la société à un rythme sans précédent, suscitant autant d'espoir que d'inquiétude. Des acteurs extrêmement enthousiastes n'hésitent pas à annoncer l'arrivée d'une IA générale, voire « super » (des systèmes d'IA atteignant, voire dépassant, les capacités humaines dans toute tâche cognitive) pour la ou les décennies à venir, rejoints dans leur posture prophétique par d'autres, plus alarmistes, qui mettent en garde contre le risque existentiel que cela représenterait pour l'humanité. D'une manière bien plus tangible et concrète, l'IA est largement présentée comme l'un des principaux moteurs économiques de l'époque actuelle. Il faut apprendre à surfer sur la vague qui s'annonce. Les entreprises doivent intégrer l'IA aussi profondément que possible dans toutes leurs activités pour maintenir leur compétitivité. Pour rester (ou devenir) des acteurs de premier plan, les nations et les acteurs technologiques doivent se livrer à une course effrénée et développer le plus rapidement possible les LLM et les systèmes d'IA générative les plus puissants. Toute tentative de réglementation, même légèrement contraignante, suscite son lot de protestations, présentant l'effort réglementaire comme un suicide économique. Comme l'affirme le sociologue français Dominique Boullier, nos sociétés sont prises au piège dans un « tunnel cognitif » marqué par une sorte de « **tyrannie de la lenteur** ».[1]

Une atmosphère sociétale aussi agitée ne constitue pas un cadre idéal pour le développement et l'utilisation éthiques de l'IA, d'autant plus qu'elle s'inscrit dans un contexte **d'asymétries de pouvoir croissantes** en faveur des géants de la tech et des cercles financiers associés. Des études récentes ont montré comment certains grands acteurs et investisseurs technologiques américains ont acquis une influence croissante sur les processus démocratiques et de gouvernance, notamment en raison de leur position privilégiée dans l'orientation éditoriale de nos paysages informationnels, mais aussi de leur implication en tant que sous-traitants ou fournisseurs de solutions dans des domaines publics nationaux tels que la sécurité.[2] Certains de ces acteurs promeuvent ou adhèrent à des idéologies techno-solutionnistes, comme l'illustre de manière frappante le Manifeste techno-optimiste du capital-risqueur Marc Andreessen (« Nous croyons qu'il n'existe aucun problème matériel – qu'il soit créé par la

[1] Dominique Boullier, « Sommet IA : la nécessaire sécession sémantique européenne - AOC media », *AOC media - Analyse Opinion Critique*, 9 février 2025 <https://aoc.media/analyse/2025/02/09/sommet-ia-la-necessaire-secession-semantique-europeenne/> [consulté le 21 octobre 2025].

[2] Shoshana Zuboff, *L'ère du capitalisme de surveillance : la lutte pour un avenir humain à la nouvelle frontière du pouvoir* (PublicAffairs, 2020) ; Rob Lalka, *Les alchimistes de la start-up : comment les géants de la tech ont transformé les profits en pouvoir* (Columbia University Press, 2024) ; Marietje Schaake, *Le coup d'État technologique : comment sauver la démocratie de la Silicon Valley* (Princeton University Press, 2024).

nature ou par la technologie – qui ne puisse être résolu par davantage de technologie »).[3] Comme l'a longtemps souligné Hans Jonas, une telle relation d' e non critique à la technologie peut permettre au développement technologique de « [prendre] continuellement un nouvel élan, entraînant ses porteurs comme des instruments désignés ».[4]

Dans ce contexte, la gouvernance démocratique du développement technologique en général et de l'IA en particulier semble plus indispensable que jamais (malgré sa nature complexe, notamment au regard des impératifs économiques). Ce besoin est bien perçu par l'opinion publique.[5] Comme l'a déclaré Zuboff en résumant l'esprit qu'elle perçoit dans la loi européenne sur les services numériques, « **le numérique doit vivre dans la maison de la démocratie** ».[6] Cela étant dit, il est important de préciser que le défi de la régulation démocratique ne se réduit pas aux déséquilibres de pouvoir. Les difficultés découlent également du rythme et des incertitudes du développement de l'IA et de ses conséquences, générant par exemple ce qu'on appelle le « dilemme de la preuve », avec la nécessité de trouver un équilibre entre des « mesures préventives excessives d'atténuation des risques fondées sur des preuves limitées » et le danger d'« attendre des preuves plus solides d'un risque imminent ».[7] De plus, les modalités de la réglementation soulèvent de multiples questions : quel équilibre entre instruments juridiquement contraignants et outils de droit souple ? Quels principes et valeurs devraient guider la réglementation ? Qu'en est-il de la participation des citoyens et des parties prenantes ? Sous quelles formes ? Comment garantir l'application des réglementations adoptées ?

Au-delà d'un simple défi réglementaire, ce à quoi nos sociétés sont confrontées, c'est la nécessité de s'engager dans un « travail de réinvention attendu depuis longtemps ».[8] Au cours de ses trois premières années d'activité (2022-2025), les membres du réseau NHNAI ont mis leurs ressources au service d'un tel effort collectif d'exploration, notamment à travers le prisme de l'humanisme et de la question de ce que signifie être humain à l'ère de l'IA. S'appuyant sur les conclusions et les enseignements tirés de cette première phase de fonctionnement, ce livre blanc vise à proposer des recommandations pour relever le défi consistant à favoriser le développement et les utilisations de systèmes d'IA qui seraient véritablement au service de l'humanité. Ces recommandations s'articuleront autour de trois axes principaux : 1) des

---

[3] Marc Andreessen, « The Techno-Optimist Manifesto », *Andreessen Horowitz*, 16 octobre 2023 <https://a16z.com/the-techno-optimist-manifesto/> [consulté le 24 juin 2025].

[4] Hans Jonas, *Essais philosophiques : des croyances antiques à l'homme technologique* (Prentice-Hall, 1974), p. 48.

[5] « Il existe un fort mandat public en faveur de la réglementation de l'IA, 70 % des personnes interrogées estimant qu'une réglementation est nécessaire. Cependant, seules 43 % d'entre elles jugent les lois actuelles adéquates. Les gens attendent des lois internationales (76 %), une réglementation nationale (69 %) et une corégulation avec l'industrie (71 %). 87 % souhaitent également des lois et des vérifications des faits pour lutter contre la désinformation générée par l'IA. » Nicole Gillespie et al., *Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025* (Université de Melbourne et KPMG, 2025), p. 5, doi:10.26188/28822919.

[6] Shoshana Zuboff, « La démocratie peut encore mettre fin à la domination des géants de la tech sur nos vies », TIME, 5 mai 2022 <https://time.com/6173639/democracy-big-techs-dominance-shoshana-zuboff/> [consulté le 25 octobre 2025].

[7] Yoshua Bengio, *Rapport international sur la sécurité de l'IA* (2025), p. 14.

[8] Zuboff, « La démocratie peut encore mettre fin à la domination des géants de la tech sur nos vies ».

recommandations sur la manière dont la réflexion de fond visant à soutenir la réglementation devrait être conçue et organisée, 2) des recommandations sur les contenus fondamentaux susceptibles de constituer le socle de cette réflexion de fond (éléments sans lesquels la réflexion pourrait être compromise), 3) des recommandations sur les thèmes importants qui devraient être explorés dans le cadre de cette réflexion de fond.

## Quels efforts de fond pour soutenir une réglementation démocratique forte ?

Il est désormais largement reconnu que la réglementation de l'IA nécessite la contribution des acteurs de la société civile. La question importante est alors de déterminer quels types d'efforts et de réflexions doivent être mis en place pour produire ces contributions. Il est particulièrement important de discuter de qui doit être impliqué, de ce que chacun doit faire et dans quel but.

### Une réflexion inclusive et horizontale (recommandation 1)

En ce qui concerne la première question, il est important de souligner les limites de tout cadre technocratique ou élitiste. On pourrait par exemple être tenté de déléguer l'effort de conception de la réglementation uniquement à des groupes d'experts (issus des domaines scientifiques et techniques ainsi que des sciences humaines et sociales, etc.) et à des représentants politiques qui adopteraient une position de surplomb épistémique. Cependant, rien ne garantit que ces groupes, bien qu'indispensables, puissent à eux seuls parvenir à des réponses correctes. Il semble important de déployer des efforts plus inclusifs, en mobilisant toutes les parties prenantes concernées. Indépendamment de cette nécessité d'impliquer les parties prenantes quant à la qualité du contenu de la réglementation elle-même, une participation sociale trop faible conduirait à une faible sensibilisation de la société à l'existence même et au contenu des dispositifs réglementaires, alors qu'il s'agit d'un élément clé d'une application rigoureuse des réglementations.[9]

### Une réflexion non seulement sur les utilisations des systèmes d'IA, mais aussi sur leur conception (recommandation 2)

Une question encore plus importante réside dans les objectifs, les finalités et les modalités de cette participation pour concevoir les contributions sociétales visant à réglementer l'IA. Bien que déjà bien connu, un point mérite d'être rappelé ici : **la gouvernance de la technologie (et les contributions sociétales ainsi requises) ne se réduit pas à la réglementation des utilisations**. Une discussion un peu plus approfondie pourrait s'avérer utile, car cette affirmation, aussi évidente qu'elle puisse paraître à première vue, entre en conflit avec notre tradition culturelle (marquée par la dichotomie nature-culture et la tendance à concevoir l'éthique comme une activité de délibération individuelle sur ses propres choix et actions). En

[9] Une récente étude mondiale menée par l'université de Melbourne et KPMG a démontré que « la plupart des gens ignorent l'existence des lois, de la législation ou des politiques gouvernementales qui s'appliquent à l'IA ». Voir : Gillespie et al., *Trust, Attitudes and Use of Artificial Intelligence*, p. 8.

effet, il faut résister à la vision traditionnelle (en partie héritée de la modernité) selon laquelle, d'une part, le développement technologique et l'activité scientifique sur laquelle il s'appuie relèvent du domaine des *faits*, produisant de manière neutre des vérités sur ce qui *est* et ce que cela signifie pour l'action ; et d'autre part, la société (la volonté individuelle et démocratique) règne sur le domaine des *valeurs*, décidant librement de ce *qu'il convient* de faire de ces connaissances et de ces pouvoirs. Selon cette dichotomie faits-valeurs, la volonté sociétale (individuelle ou démocratique) ne fait que compléter le tableau en choisissant ce qu'il convient de faire des faits neutres et des moyens technologiques. Cependant, cette vision est bien trop simpliste. De nombreux critiques ont remis en cause la possibilité même d'une dichotomie faits-valeurs.[10] Assurer une régulation démocratique et sociétale forte de l'IA nécessite de bien comprendre les relations subtiles et **les imbrications entre technologie et éthique** (comprise comme la réflexion, la délibération et l'action dans le domaine de ce qui *devrait être*).

Comme l'a clairement montré la tradition de longue date de la critique technologique, les artefacts comportent presque toujours des implications politiques et éthiques.[11] Ainsi, la plupart du temps, **l'éthique intervient bien en amont de l'utilisation**, dès la phase de conception des technologies. Cela vaut tout particulièrement pour les technologies d'IA, qui disposent souvent d'une plus grande autonomie que les artefacts plus traditionnels. Les inquiétudes concernant la perte potentielle de contrôle sur de puissants systèmes d'IA générative en constituent une illustration extrême,[12] évoquant des catastrophes sans qu'il y ait de choix d'utilisation en soi, simplement parce que la technologie existe. Aussi frappantes et inquiétantes que puissent être ces perspectives, nous ne devons pas les laisser occulter le fait que la nécessité d'une réflexion éthique en amont s'applique de manière bien plus généralisée.

Comme l'a proposé Moor en 2006,[13] nous devrions considérer la plupart des dispositifs numériques comme des « agents à impact éthique », dans la mesure où ils produisent des résultats présentant une pertinence éthique. Cette manière d'aborder la question montre clairement jusqu'où l'éthique peut remonter en amont. La réflexion éthique est en quelque sorte présente dès l'instant où l'on tente de construire un *bon* artefact. En un sens, les objectifs mêmes de fiabilité et de sécurité de nos ordinateurs et programmes constituent déjà des enjeux éthiques auxquels les ingénieurs sont chargés de répondre à travers la conception des artefacts. Les fonctionnalités garantissant différents degrés de confidentialité correspondent à

---

[10] Voir par exemple : Hilary Putnam, *The Collapse of the Fact-Value Dichotomy, and Other Essays Including the Rosenthal Lectures* (Harvard University Press, 2002).

[11] Langdon Winner, « Do Artifacts Have Politics? », *Daedalus*, 109.1 (1980), pp. 121–36.

[12] Une lettre ouverte récente, qui a reçu le soutien de nombreuses personnalités publiques et de chercheurs renommés, exprime de profondes inquiétudes liées à l'émergence possible d'une super IA, « allant de l'obsolescence économique et de la perte d'autonomie des humains, à la perte de liberté, des libertés civiles, de la dignité et du contrôle, en passant par des risques pour la sécurité nationale et même l'extinction potentielle de l'humanité », voir : « Déclaration sur la superintelligence », Déclaration sur la superintelligence, 2025 <https://superintelligence-statement.org> [consulté le 26 octobre 2025]. Pour une analyse moins médiatisée mais plus approfondie, voir : Bengio, *Rapport international sur la sécurité de l'IA*, sec. 2.2.3.

[13] J.H. Moor, « The Nature, Importance, and Difficulty of Machine Ethics », *IEEE Intelligent Systems*, 21.4 (2006), pp. 18–21, doi:10.1109/MIS.2006.80.

des choix éthiques de conception auxquels nous sommes plus habitués. Pour en revenir à l'IA, on pourrait mentionner les cas bien étudiés d'outils de classification biaisés ou de systèmes de décision algorithmiques[14] (en particulier lorsqu'ils sont mobilisés pour l'administration des services publics), ou encore d'algorithmes de recommandation qui éditorialisent l'énorme quantité d'informations disponibles sur nos marchés libéralisés de l'information numérique (sur Internet ou les plateformes sociales) à des fins de captation de l'attention.[15] Là encore, la « qualité » éthique des résultats ne dépend pas uniquement (ni même principalement) des choix opérés au niveau des utilisations. **La qualité éthique est en grande partie intégrée dès la conception des systèmes**.[16]

Sur la base de ce rappel conceptuel structurant, il est donc clair que la réglementation et la réflexion éthique doivent également intervenir au stade précoce du développement des systèmes d'IA et pas seulement au niveau de l' , lors de l'utilisation de solutions toutes faites. Comme le formule Dignum, une « IA responsable » nécessite **une éthique pour, dans et par la conception**.[17] Et là encore, la participation des citoyens et des parties prenantes est nécessaire pour alimenter la régulation externe des processus de développement des systèmes d'IA, mais aussi au sein même de ces processus, par le biais d'approches de R&D dédiées telles que la conception sensible à la valeur (VSD), la recherche et l'innovation responsables (RRI) ou la conception participative (PD).[18]

---

[14] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, première édition de poche (B/D/W/Y Broadway Books, 2017) ; David Restrepo Amariles, « Algorithmic Decision Systems: Automation and Machine Learning in the Public Administration », dans *The Cambridge Handbook of the Law of Algorithms*, édité par Woodrow Barfield, Cambridge Law Handbooks (Cambridge University Press, 2020), pp. 273–300, doi:10.1017/9781108680844.015.

[15] Gérald Bronner, *Apocalypse cognitive* (PUF, 2021).

[16] Moor propose une distinction supplémentaire utile. Lorsque des résultats présentant un intérêt éthique sont obtenus ou générés directement en raison de la conception du système, Moor parle d'« agents éthiques implicites ». Cela s'oppose aux « agents éthiques explicites », qui manipulent explicitement des éléments éthiques (comme pourraient l'être des algorithmes de modération ou des outils visant à éliminer les biais des bases de données, qui s'appuieraient par exemple sur des règles et des critères éthiques explicites). Cette deuxième catégorie de systèmes correspond à un sous-domaine à part entière de l'IA appelé « éthique des machines ». Voir par exemple : *Machine Ethics*, édité par Michael Anderson et Susan Leigh Anderson, 1re éd. (Cambridge University Press, 2011), doi:10.1017/CBO9780511978036.

[17] L'éthique dans, par et pour la conception correspond respectivement aux niveaux implicite et explicite évoqués par Moor, ainsi qu'au contexte éthique entourant le travail de conception (pression économique, équilibre entre les sexes dans les équipes d'ingénierie logicielle, par exemple). Voir : Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Artificial Intelligence: Foundations, Theory, and Algorithms (Springer International Publishing, 2019), doi:10.1007/978-3-030-30371-6.

[18] Par exemple, le groupe d'experts de haut niveau de l'UE sur l'IA recommande de mobiliser des méthodes techniques et non techniques, en s'appuyant sur la « participation des parties prenantes et le dialogue social », afin de mettre en place une « IA *digne de confiance* tout au long du cycle de vie du système » ; voir :(2019), pp. 20–23 <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [consulté le 8 août 2025]. (p. 20-23). Pour une discussion plus académique et des détails sur les méthodes participatives, voir : Dignum, *Responsible Artificial Intelligence* ; Till Winkler et Sarah Spiekermann, « Twenty Years of Value Sensitive Design: A Review of Methodological Practices in VSD

Un effort de transformation fondé sur une véritable réflexion politique, avec le soutien d'experts (recommandation 3)

[19]En effet, **la conception des systèmes d'IA** comporte une **dimension politique** inéluctable, une dimension qui doit être reconnue dans toute son ampleur. En particulier, les contributions sociétales et la participation des parties prenantes ou des citoyens ne peuvent se réduire à une simple collecte neutre des mentalités par le biais de sondages d'opinion ou d'enquêtes sociologiques. On peut difficilement supposer qu'elles fourniront toujours des contributions exploitables, prêtes à l'emploi et légitimes pour guider les processus de conception. Tout d'abord, s'assurer d'inclure toutes les voix légitimes et de représenter correctement ce qu'elles expriment peut s'avérer extrêmement difficile. De plus, les préférences, les valeurs et les principes éthiques auxquels adhèrent les personnes peuvent entrer en conflit (la question de la surveillance assistée par la reconnaissance faciale IA illustre bien ces tensions, avec le dilemme entre les objectifs d'amélioration de la sécurité et de préservation de la vie privée et de la liberté). Bien qu'extrêmement précieuses, ces simples instantanés de l'opinion publique ne peuvent être rien de plus que des points de départ.

Dans bon nombre des questions éthiques profondes soulevées par l'IA, l'état actuel de la société et les préférences des individus font davantage partie du problème que de la solution. Les algorithmes de classification biaisés le sont en grande partie à cause des exemples fournis au départ par la société (nous reviendrons plus en détail sur ce point dans la section suivante). Dans le même ordre d'idées, la menace que les algorithmes de recommandation font peser sur notre capacité à faire bon usage de notre temps d'attention libre est en grande partie due au modèle économique gratuit (adopté sans discernement par de nombreux clients). Ainsi, résoudre ces problèmes ne signifie pas seulement orienter correctement la conception et les utilisations de l'IA, mais nécessite également des transformations sociales pour rendre cette orientation possible. De plus, ces efforts de transformation nécessiteront souvent l'apport d'experts issus de nombreuses disciplines différentes (philosophes, économistes, psychologues, historiens, informaticiens…) pour soutenir les participants dans leur réflexion. À cet égard, et comme l'illustrent bien les exemples ci-dessus, de nombreuses questions éthiques liées à l'IA peuvent être considérées comme des « problèmes **épineux** », des problèmes sociétaux complexes « tels que la violence, la faim, la pauvreté, la maladie et la pollution environnementale », qui relèvent autant des domaines techniques et scientifiques (y compris les sciences humaines) que du domaine politique.[20] Ils constituent des interrogations et des

Projects », *Ethics and Information Technology*, 23.1 (2021), pp. 17–21, doi:10.1007/s10676-018-9476-2 ; Carolyn Ten Holter, « Participatory Design: Lessons and Directions for Responsible Research and Innovation », *Journal of Responsible Innovation*, 9.2 (2022), pp. 275–90.

[19] Comme le rapport déjà mentionné de l'Université de Melbourne et de KPMG qui « vise à fournir une compréhension fondée sur des données factuelles de la confiance, de l'utilisation et des attitudes des personnes à l'égard de l'IA, de leurs points de vue sur les impacts de l'IA, et de leurs attentes en matière de gouvernance et de réglementation » ( Gillespie et al., *Trust, Attitudes and Use of Artificial Intelligence*, p. 4.

[20] Christian Pohl, Bernhard Truffer et Gertrude Hirsch-Hadorn, « Addressing Wicked Problems through Transdisciplinary Research », dans *The Oxford Handbook of Interdisciplinarity*, édité par Robert Frodeman (Oxford University Press, 2017), p. 0, doi:10.1093/oxfordhb/9780198733522.013.26.

défis politiques et éthiques profonds sur la manière dont les sociétés devraient être organisées et dont les individus devraient vivre, tout en intégrant des questions techniques et scientifiques.

En conséquence, l'élaboration de contributions sociétales susceptibles d'orienter la réglementation et la conception de l'IA exige souvent une coproduction « transdisciplinaire » de solutions,[21] où des spécialistes de nombreuses disciplines différentes mettent leur expertise au service de réflexions collectives dans lesquelles les citoyens et les parties prenantes sont les acteurs centraux. En somme, la participation doit inclure, avec l'aide d'experts, des processus **d'autocritique, d'enrichissement mutuel et de transformation des groupes sociaux eux-mêmes**.[22] Une fois encore, il convient de distinguer cela de toute approche technocratique dans laquelle des experts, depuis une position épistémique dominante, dicteraient aux acteurs de la société ce qu'ils doivent faire. Les experts n'apportent que quelques pierres pour l'édifice que le collectif doit construire. Ils soutiennent la réflexivité, l'autocritique et l'autotransformation du collectif et en tirent eux-mêmes profit (en tant que chercheurs, mais aussi en tant que citoyens).

Affirmer, comme nous venons de le faire, que les groupes sociaux doivent s'engager dans des processus réflexifs et transformateurs afin d'apporter une contribution *plus pertinente* à la réglementation et au développement de l'IA peut sembler assez problématique. Cela entre en conflit avec la tradition culturelle que nous avons mentionnée plus haut. En effet, la dichotomie entre faits et valeurs est souvent invoquée pour justifier l'idée que – contrairement aux questions factuelles qui peuvent être étudiées de manière empirique ou scientifique, et donc tranchées de manière convaincante – les questions éthiques, politiques et, plus largement, évaluatives n'appartiennent pas au domaine de la connaissance, de la rationalité et de la vérité, mais à celui de la liberté pure et du libre arbitre. Par conséquent, les individus sont libres d'adopter la vision qu'ils souhaitent quant à ce qui *devrait être*. Et si n'importe qui d'autre est totalement libre de ne pas être d'accord, personne ne peut dire à quelqu'un d'autre qu'il a tort (lui dire que ses affirmations sont *fausses*). Nous acceptons d'être en désaccord. Nous n'avons tout simplement pas les mêmes valeurs. Tout au plus, nous pouvons essayer de rallier quelqu'un d'autre à nos intérêts pour des raisons purement pragmatiques (à la manière des sophistes), mais pas dans la logique d'une quête collective de la vérité, d'une amélioration des points de vue partagés (ce dont nous avons précisément tenté de défendre la nécessité). Là encore, la dichotomie faits-valeurs est à la fois erronée et dangereusement trompeuse. Premièrement, la philosophie des sciences et l'épistémologie de la seconde moitié du XXe siècle ont clairement montré que les investigations scientifiques et factuelles mobilisent de manière irréductible des jugements d' s évaluatives.[23] De plus, il est (évidemment ?) loin d'être évident qu'il ne puisse y avoir de connaissance ou de rationalité évaluative (y compris éthique et

[21] Julie Thompson Klein, « Typologies of Interdisciplinarity: The Boundary Work of Definition », dans *The Oxford Handbook of Interdisciplinarity*, éd. par Robert Frodeman, 2e éd. (Oxford University Press, 2017), pp. 21–34 (section 3.5), doi:10.1093/oxfordhb/9780198733522.013.3.

[22] Florin Popa, Mathieu Guillermin et Tom Dedeurwaerdere, « Une approche pragmatique de la transdisciplinarité dans la recherche sur la durabilité : de la théorie des systèmes complexes à la science réflexive », *Futures*, 65 (2015), p. 45-56, doi:10.1016/j.futures.2014.02.002.

[23] Julian Reiss et Jan Sprenger, « Scientific Objectivity », dans *The Stanford Encyclopedia of Philosophy*, éd. par Edward N. Zalta, hiver 2020 (Metaphysics Research Lab, Université de Stanford, 2020) <https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/> [consulté le 27 mai 2025].

morale).[24] Il est donc légitime de parler de ***meilleures* contributions sociétales au sens fort** (et non pas seulement meilleures par rapport à certains groupes ou intérêts particuliers).

Bien sûr, notre intention ici n'est pas de contester la légitimité de la liberté démocratique et du pluralisme démocratique. Ce à quoi nous pensons qu'il faut s'opposer, c'est la tendance à les caricaturer en visions relativistes ou solipsistes qui dissocient cette liberté de la responsabilité et du devoir de rechercher sincèrement la vérité ou la validité dans tout ce que nous pensons, affirmons et faisons. Comme l'explique bien Revault d'Allonnes, ce découplage est à l'origine des abus de l'ère post-vérité et, bien qu'il puisse donner l'illusion d'une liberté accrue, il sape en réalité la capacité des individus à véritablement habiter leur monde (car ils deviennent incapables de reconnaître les preuves factuelles concernant la manière dont le monde est, ce qui est une condition préalable pour imaginer d'autres façons dont il pourrait être et commencer à le transformer).[25] Appliqué à nos thèmes concernant l'élaboration de contributions sociétales pour guider la réglementation, le développement et l'utilisation de l'IA, nous ne pouvons nous contenter (au nom d'un pluralisme démocratique abstrait) d'un simple patchwork d'opinions préconçues plus ou moins divergentes (chaque acteur se contentant de tolérer les opinions contradictoires et, par exemple, de se plier à la volonté de la majorité), ni ne sommes condamnés à nous en tenir là. Nous avons la possibilité, et donc le devoir, de mettre en place **des communautés de discussion et de délibération rationnelles**[26] qui s'attachent à élaborer une compréhension et des orientations éthiques concernant les enjeux sociétaux de l'IA. En particulier, au lieu de traiter systématiquement les revendications divergentes comme des questions d'opinions individuelles libres, nous pouvons tenter de parvenir à un accord de base sur le fait qu'au moins certaines divergences sont le signe de **tensions** légitimes **ou de complexités inhérentes à la question explorée**. Prenons par exemple les objectifs de sécurité et de protection de la vie privée ou des droits fondamentaux, qui s'opposent lorsqu'il s'agit de la surveillance assistée par l'IA. Bien que les opinions puissent diverger quant à l'ordre de priorité à accorder à l'un ou l'autre, il semble possible que tous s'accordent sur la légitimité des deux objectifs. Déployer des efforts communs pour établir rationnellement un tel terrain d'entente améliorerait sans aucun doute la réflexion éthique collective.

### Renforcement des capacités éthiques pour des forces motrices ascendantes et horizontales (recommandation 4)

Dans un certain sens, les efforts réflexifs et transformationnels nécessaires pour produire des contributions sociétales susceptibles d'orienter le développement et la réglementation de l'IA peuvent être considérés comme des processus de renforcement des capacités éthiques. En déployant un travail sincère et collectif pour cartographier les enjeux éthiques et leurs complexités et pour tenter d'élaborer des réponses, les communautés participantes cultiveront

[24] Aaron Zimmerman, *Moral Epistemology*, New Problems of Philosophy (Routledge, 2010), doi:10.4324/9780203850862.

[25] Myriam Revault d'Allonnes, *La faiblesse du vrai : ce que la post-vérité fait à notre monde commun*, La couleur des idées (Éditions du Seuil, 2018).

[26] Philip Kitcher, *Science, Truth and Democracy*, Oxford Studies in the Philosophy of Science (Oxford University Press, 2001).

leurs capacités à contribuer à l'orientation éthique de l'IA. Une telle autonomisation est indispensable pour de multiples raisons. Tout d'abord, il est nécessaire que les communautés de citoyens et de parties prenantes aident l' e à concevoir des principes éthiques généraux et des réglementations juridiques. Mais au-delà de cette dimension globale de la réglementation, la présence de communautés dotées de capacités éthiques renforcées est également cruciale pour contextualiser et appliquer concrètement les principes éthiques ou juridiques généraux dans les contextes locaux. Cela peut considérablement faciliter l'application et le respect des réglementations juridiques (la prise de conscience et les attentes ascendantes locales venant à la rencontre des efforts réglementaires descendants). Les communautés locales peuvent également contribuer à rendre plus efficaces les dispositifs de soft law et autres types de régulation plus horizontaux et non contraignants juridiquement (labels, chartes, lignes directrices, etc.). Des choix éclairés de la part des consommateurs seront également essentiels, ouvrant des espaces économiques viables pour un entrepreneuriat plus éthique.

De plus, et c'est peut-être important, des communautés autonomisées peuvent participer à l'identification des besoins et à la définition de cas d'utilisation à forte valeur ajoutée des technologies disponibles. Nous développerons certains aspects de cette question un peu plus loin, mais nous devons déjà souligner ici le danger qu'il y aurait à croire en une sorte d'utilité par principe des technologies d'IA (en particulier de l'intelligence artificielle générale ou super-IA tant prophétisée). Divers éléments technologiques formidables sont disponibles, et il ne fait aucun doute que de nouveaux seront développés. Cependant, trouver des cas d'utilisation véritablement positifs pour ces technologies pourrait constituer l'un des défis les plus sérieux de l'éthique et de la réglementation de l'IA. Une contribution élargie et éclairée de toutes les parties prenantes concernées à ce stade crucial du développement de l'IA sera décisive pour que les technologies d'IA contribuent véritablement aux sociétés humaines dans leurs différentes dimensions. En résumé, un renforcement des capacités éthiques aussi large que possible est nécessaire pour favoriser la réglementation **ainsi que** le développement et l'utilisation éthiques de l'IA en créant ou en renforçant **des forces d'orientation descendantes, ascendantes et horizontales**.

## Une réflexion approfondie sur ce que signifie être humain (recommandation 5)

Penchons-nous maintenant sur le contenu, les thèmes sur lesquels ces efforts de renforcement des capacités éthiques pourraient porter. À cet égard, il est particulièrement important de replacer les questions générales dans le bon ordre de causalité. On entend trop souvent des injonctions invitant les humains à trouver leur place dans le nouveau monde façonné par l'IA. Joshua Bengio et les autres coauteurs de *l'International AI Safety Report* (2025) l'ont exprimé très clairement (et nous y reviendrons dans la section suivante) : « L'IA ne nous arrive pas : ce sont les choix faits par les gens qui déterminent son avenir ».[27] Cela signifie que la bonne question est « quelle est la place de l'IA dans notre monde humain et notre environnement naturel ? ». Idéalement, nous devrions orienter l'IA en réfléchissant profondément à nos besoins, ce qui implique à son tour un effort mondial pour construire une compréhension commune des projets de société que l'IA devrait servir.

[27] Bengio, *International AI Safety Report*, p. 14.

Dans ce qui suit, nous souhaitons nous concentrer sur une dimension importante de ces questions. En effet, pour y répondre, il faut également réfléchir à ce que signifie « être humain », à ce que nous sommes, mais surtout à ce que nous voulons ou devons être en tant qu'êtres humains. Cela nous amène au thème de l'humanisme dans ses différentes acceptions (non seulement descriptive, historique ou culturelle, mais aussi axiologique). De nombreux principes majeurs de l'éthique de l'IA font appel à l'idée de l'humain : l'IA doit être centrée sur l'humain, au service de l'épanouissement *humain*, *l'humain* doit être maintenu dans la boucle, etc. Plus fondamentalement, on peut même percevoir un lien presque organique entre l'éthique et la quête visant à comprendre ce que signifie être humain. Ricoeur, par exemple, définit la « finalité éthique » comme « l' e de la « bonne vie » avec et pour les autres dans des institutions justes ».[28] Comment pouvons-nous commencer à forger le sens de termes tels que « bonne vie » ou « institutions justes » sans réfléchir en même temps à ce que signifie être humain ? De plus, les nouvelles technologies telles que l'IA (mais on pourrait également inclure la neurotechnologie dans ce tableau) offrent aux sciences humaines des capacités accrues d'auto-modification en profondeur (une *anthropotechnè* dans le cadre de laquelle les humains peuvent influencer leur propre nature).[29] Comment devrions-nous nous transformer, dans quelle direction ? Une boussole fiable concernant l'humanisme et ce que signifie être humain semble plus nécessaire que jamais.

Et c'est bien un effort d'exploration et d'approfondissement qui s'impose. La notion d'humanisme est loin d'être claire et consensuelle, sans problème et prête à servir d'orientation éthique. Même considérée uniquement dans le contexte de son émergence (en Europe avec la Renaissance, puis la modernité et les Lumières), la notion intègre déjà des dimensions multiples, disparates et parfois conflictuelles.[30] Par la suite, de nombreux courants se sont opposés et continuent de s'opposer de front à l'humanisme de la modernité (anti-, post-, transhumanisme), en soulignant ses difficultés et ses limites. Existe-t-il des caractéristiques précises qui distinguent l'humain du non-humain ? Sont-elles universelles ? La primauté quasi absolue accordée par l'humanisme moderne à l'autonomie et à la rationalité humaines n'est-elle pas problématique ? N'a-t-elle pas conduit au mythe de l'humain en tant que maître et possesseur de la nature, avec un lien automatique entre développement technoscientifique et progrès humain ? Aujourd'hui, ces difficultés semblent être renforcées par les contributions théoriques et scientifiques de l'IA et des sciences neuves, ainsi que par leurs retombées technologiques.

Mais peut-être la solution ne consiste-t-elle pas à rejeter d'emblée la notion d'humanisme. Il est également possible de remettre cette notion au travail, à travers une exploration renouvelée

---

[28] Paul Ricoeur, *Soi-même comme un autre*, *L'Ordre philosophique* (Éditions du Seuil, 1990), p. 202, notre traduction : « visée éthique » est définie comme « la visée de la "vie bonne" avec et pour autrui dans des institutions justes ».

[29] Sylvain Lavelle, « What a Human Is, Could Be and Should Be. The Anthropology of the Human and the Philosophy of Humanism », dans *Human Freedom at the Test of AI and Neuroscience*, édité par Stefano Biancu, Mathieu Guillermin et Fabio Macioce, Contemporary Humanism: Open Access Annals (2024) (Edizioni Studium, 2024), pp. 119–41 <https://www.edizionistudium.it/riviste/studium-contemporary-humanism-open-access-annals-2024>.

[30] Voir par exemple la présentation historique et philosophique éclairante de la notion d'humanisme : Tony Davies, *Humanism* (Routledge, Taylor & Francis Group, 2001).

capable de préserver et d'approfondir les contributions les plus fructueuses de l'humanisme de la modernité et des Lumières. Il est courant, par exemple, de ne retenir de Kant que l'idée d'une raison humaine capable de parvenir à des conclusions *a priori* dans les domaines de la science (sur les phénomènes) et de la morale (avec l'impératif catégorique) – la raison pure théorique et pratique. Il est donc trop courant de pointer du doigt l'échec de l'épistémologie kantienne (par exemple, le dépassement de la physique newtonienne) et de discréditer la modernité et l'humanisme. Mais ce serait ignorer le cœur du kantisme, avec son idée centrale du passage de l'humanité de la minorité à la majorité (l'individu peut et doit penser par lui-même, « *Sapere* Aude ») et l'importance du couple fondamental liberté-responsabilité, qui conduit Kant à placer la raison pratique, et l' e surtout la faculté de jugement, au centre de tout son système philosophique.[31] Une modernité « critique » plutôt que dogmatique est envisageable.[32]

Se priver de ce type d'apport ne ferait qu'alimenter les difficultés liées à l'éthique, à la gouvernance et à la réglementation de l'IA. Il semble bien plus fructueux d'approfondir ces ressources et de les associer aux perspectives passionnantes de l'IA, des neurosciences et des sciences cognitives afin de dessiner les contours d'un nouvel humanisme, s'ouvrant à une compréhension renouvelée de notre liberté, de notre intelligence ou de notre capacité de jugement. Pour soutenir l'exploration de cette question de ce que signifie être humain, il peut d'abord être utile de dresser un tableau des étapes fondamentales concernant la réalité de la technologie de l'IA et, par là même, la possibilité de mettre en évidence certaines spécificités des humains par comparaison avec les machines.

## Étapes fondamentales concernant l'IA et les humains

Pour explorer, construire et réfléchir aux projets de société auxquels l'IA devrait servir, sur la base d'une exploration collective approfondie de ce que signifie (ou devrait signifier) être humain, une condition préalable est de garantir une compréhension solide et enrichissante de ce qu'est l'IA. Malheureusement, comme évoqué dans l'introduction, l'IA est souvent mal représentée ou dépeinte de manière trop abstraite. Il est par exemple courant de présenter les divers succès que l'apprentissage (profond) automatique a permis de produire comme les succès **d'une grande entité unique appelée IA**. En particulier, il est très impressionnant d'imaginer cette grande IA comme, d'une part, capable de produire de nouvelles (et peut-être meilleures) façons d'accomplir une tâche (AlphaGo et la découverte d'une nouvelle manière puissante de jouer au Go, par exemple) et, d'autre part, capable de répondre à toutes nos questions, en particulier les plus difficiles (comme avec les versions récentes de GPT qui peuvent répondre à des questions de niveau doctorat en sciences naturelles).

[31] Alexis Philonenko, *L'œuvre de Kant : la philosophie critique. 1 : La philosophie pré-critique et la critique de la raison pure*, À la recherche de la vérité, 6e éd. (Vrin, 1996) ; Alexis Philonenko, *L'œuvre de Kant : la philosophie critique. 2 : Morale et politique*, À la recherche de la vérité, 5e éd. (Vrin, 1997).

[32] Bernard Feltz, *La science et le vivant : philosophie des sciences et modernité critique*, 2e éd. revue et augmentée (De Boeck, 2014).

Cette façon de présenter les choses donne à penser qu'avec suffisamment de données et de puissance de calcul, l'IA n'a pratiquement aucune limite. Sur cette base, nombreux sont ceux qui annoncent l'émergence prochaine d'une intelligence artificielle générale (AGI) voire d'une super-intelligence artificielle (SAI), alimentant ainsi l'idée d'une compétition (en fin de compte perdue d'avance) entre les humains et l'IA dans le domaine de l'intelligence. Les croyances sous-jacentes à ces questions sont souvent quasi-religieuses, comme l'illustrent les récentes déclarations d'Eric Schmidt : la SAI est telle qu'elle « peut prouver quelque chose que nous savons être vrai, mais dont nous ne pouvons pas comprendre la preuve ».[33] Ce type d'atmosphère alimente également la tendance à présenter le problème de la perte potentielle de contrôle sur les systèmes d'IA en termes de conflit ou de guerre entre les intelligences. Même Geoffrey Hinton, lauréat du prix Nobel de physique 2024 (pour ses recherches fondamentales qui ont ouvert la voie aux techniques modernes d'apprentissage automatique), a affirmé que le problème du contrôle est crucial car « il existe très peu d'exemples où des entités plus intelligentes sont contrôlées par des entités moins intelligentes » (l'un des rares exemples étant une mère contrôlée par son bébé).[34] De telles représentations de **l'IA comme une sorte de vague inexorable qui rendra les humains cognitivement obsolètes** sont particulièrement néfastes, d'autant plus **qu'elles occultent complètement la dimension politique et éthique de l'IA en tant qu'artefact humain**. Leur nocivité est renforcée par le niveau souvent faible de connaissances des profanes sur les technologies actuelles d'IA.[35] En conséquence, il est essentiel de garantir une compréhension pas trop abstraite de l'IA au sein des communautés engagées dans l'effort de production de ressources et de conseils pour son développement et son utilisation.

Garantir une compréhension solide et responsabilisante des technologies d'IA (recommandation 6)

Voici quelques contributions à une représentation plus valorisante de l'IA.

**Démystifier l'apprentissage automatique**

L'apprentissage automatique est une notion et un domaine techniques souvent simplifiés à l'extrême pour le grand public, par exemple en affirmant qu'il s'agit de « bébés » machines ou de programmes qui apprendront d'une manière ou d'une autre à accomplir une tâche, à l'instar d'un enfant qui apprend à faire quelque chose de nouveau. De plus, l'apprentissage

[33] Eric Schmidt est l'ancien PDG de Google. Voir : Eric Schmidt, « L'IA et la genèse d'une nouvelle ère », Conférence publique, Sommet RAISE 2025, Paris, 8 juillet 2025 <https://www.youtube.com/watch?v=_gBxYL2ihc0> [consulté le 30 octobre 2025].

[34] Geoffrey Hinton, « L'intelligence numérique remplacera-t-elle l'intelligence biologique ? », Conférence publique, Romanes Lecture, Université d'Oxford, 20 février 2024 <https://www.ox.ac.uk/news/2024-02-20-romanes-lecture-godfather-ai-speaks-about-risks-artificial-intelligence> [consulté le 30 octobre 2025].

[35] Comme le rapporte une étude récente menée par un consortium de journalistes, de nombreuses personnes (ici entre 30 % et près de 50 % selon l'âge) ont tendance à considérer l'IA comme une source d'information fiable, ce qu'elle n'est pas. Plus inquiétant encore, plus d'un tiers des personnes interrogées ont tendance, lorsqu'elles reçoivent une fausse information parce que le LLM n'a pas réussi à la résumer correctement, à blâmer non seulement l'IA, mais aussi la source d'information d'origine. Voir : Union européenne de radio-télévision (UER), *News Integrity in AI assistants* (2025), p. 4 <https://www.ebu.ch/fr/research/open/report/news-integrity-in-ai-assistants> [consulté le 24 octobre 2025]. Voir également : Mark Steyvers et al., « What Large Language Models Know and What People Think They Know », *Nature Machine Intelligence*, 7.2 (2025), pp. 221–31, doi:10.1038/s42256-024-00976-7.

automatique est parfois présenté comme produisant des boîtes noires dont le fonctionnement échappe à la compréhension humaine. Et comme nous venons de le dire, ces systèmes opaques apprendraient à répondre à toutes nos questions, notamment en trouvant de meilleures solutions à bon nombre de nos problèmes. Ce type de représentation est extrêmement problématique. On **peut et** on **doit en dire beaucoup plus sur l'apprentissage automatique**. Il faut au moins expliquer que **l'apprentissage automatique revient à rechercher automatiquement une paramétrisation adéquate d'une architecture computationnelle**, dans l'espoir d'aboutir à un programme capable d'accomplir une tâche qui résistait à une programmation explicite. Par exemple, il est assez simple d'écrire un algorithme pour classer des images simples, disons des images monochromes de différentes couleurs (comparer la moyenne des valeurs codant les couleurs dans chaque pixel de chaque image suffirait). Cependant, nous ne savons pas quelles opérations l'ordinateur devrait effectuer pour classer correctement des images de multiples objets ordinaires. Dans le cas de telles tâches récalcitrantes, les techniques **d'apprentissage automatique** peuvent permettre de **contourner partiellement nos limites de programmation**.

En fait, nous ne pouvons peut-être pas prescrire étape par étape les opérations à effectuer pour classer des images d'objets du monde réel. Mais ce que nous pouvons faire, c'est sélectionner certaines classes d'opérations (par exemple des multiplications par certains coefficients, des additions et d'autres opérations mathématiques sur les nombres spécifiant les couleurs dans les pixels d'une image) et parier qu'il existe des agencements de ces opérations qui permettraient d'accomplir la tâche. Cela revient à concevoir une **architecture computationnelle avec *des paramètres libres*** (les valeurs des coefficients de multiplication, par exemple). Cette architecture d' s se décline en de nombreux programmes différents lorsque les détails de l'agencement des opérations sélectionnées sont déterminés en fixant les paramètres libres à des valeurs spécifiques. **Le pari avec l'apprentissage automatique est alors le suivant** : nous supposons (en espérant, sur la base d'une hypothèse éclairée) qu'au moins l'un de ces programmes, que nous pouvons obtenir à partir d'une paramétrisation spécifique de l'architecture que nous avons conçue, est capable d'exécuter correctement la tâche souhaitée. Nous pouvons alors écrire un programme d'essais et d'erreurs plus ou moins intelligent dont la tâche sera de tester divers ensembles de paramètres afin de trouver le plus efficace (ou du moins un suffisamment efficace). Une fois cet objectif atteint, on dit que l'on *a appris* un modèle ou un programme. Et maintenant, peut-être le plus important : **tout cela ne peut fonctionner qu'en fournissant une orientation ou un retour d'information** à ce programme de réglage des paramètres. Pour que l'apprentissage automatique soit même possible en principe, il faut concevoir un moyen de quantifier la qualité des résultats produits par un ensemble donné de paramètres.

Cette manière d'expliquer les techniques d'apprentissage automatique a le mérite de montrer clairement pourquoi l'idée d'une entité unique, puissante et gigantesque appelée IA est illusoire. Tout d'abord, il existe **des types d'architectures computationnelles** très **différents** que l'on peut essayer d'utiliser et *d'entraîner,* et qui sont plus ou moins bien adaptés à des problèmes spécifiques (réseaux neuronaux convolutifs, réseaux neuronaux récurrents, transformateurs, ... pour n'en citer que quelques-uns). De plus, **la manière dont est défini le retour d'information qui guide le processus d'apprentissage peut également varier**

**considérablement et ouvrir la voie à différents types de succès et, par conséquent, d'attentes**. Dans le cas des jeux de société tels que le Go (ainsi que pour la plupart des jeux vidéo), le retour d'information est facile à concevoir : il suffit d'utiliser le score à la fin d'une partie ou d'un match. Dans de tels cas, la paramétrisation automatique peut très bien aboutir à un programme qui trouve de nouvelles façons de maximiser le retour d'information (de nouvelles façons puissantes de jouer au Go dont les humains n'avaient pas conscience), produisant ainsi éventuellement (notamment) de merveilleux outils d'exploration assistée. Mais il s'agit là de cas de figure particuliers où le retour d'information peut être directement et explicitement formulé en termes algorithmiques ou mathématiques (même si, dans certains cas, l'algorithme est très complexe, comme dans le cas des simulateurs ou des jeux vidéo). Cependant, dans la plupart des scénarios de la vie réelle, tels que les modèles de langage à long terme (LLM) et les bots conversationnels, mais aussi la classification d'images, nous ne disposons pas d'une définition aussi directement algorithmique ou mathématique du retour d'information. Ce que nous avons à la place, ce sont des ensembles d'exemples (conversations, livres, images déjà classées...) que nous pouvons utiliser comme référence de qualité que le processus d'apprentissage tente de reproduire à travers diverses paramétrisations de l'architecture computationnelle. Mais, dans de tels cas (et ce sont les plus courants), la maximisation du retour d'information s'obtient en reproduisant les exemples. Il est donc vain d'espérer (et trompeur d'affirmer) que cela produira des résultats radicalement nouveaux que les humains eux-mêmes étaient incapables de produire auparavant (au contraire, une paramétrisation produisant de nouveaux résultats ne serait pas sélectionnée, car mal notée puisqu'elle ne reproduit pas les exemples). Au mieux, ce type d'apprentissage automatique basé sur des exemples peut conduire à des programmes dotés d'une capacité de généralisation limitée, par exemple grâce à la capacité générique que l'on pourrait tirer de la capacité à tenir des conversations crédibles.

De plus, la vision dé-abstraite de l'apprentissage automatique que nous avons proposée permet de faire la lumière sur le « sophisme de l'insondabilité », qui consiste à présenter l'apprentissage automatique (profond) comme une « » produisant des « systèmes logiciels de type boîte noire, mystérieux et inexplicables »[36] . Tout d'abord, il devrait désormais être clair que certains humains (ingénieurs ou développeurs) connaissent assez bien le type d'opérations effectuées par le logiciel (par exemple, une multitude de multiplications, d'additions et d'autres fonctions mathématiques pas si complexes dans le cas des réseaux neuronaux profonds). Cela ne signifie toutefois pas que ces programmes soient exempts de toute opacité. Pour la plupart des algorithmes d'apprentissage profond, il est par exemple très difficile de comprendre, en examinant le programme lui-même, s'il fonctionnera et pourquoi. Il y a beaucoup trop d'opérations, et elles se déroulent au niveau sous-symbolique (sur des listes de nombres représentant des objets manipulés dans des espaces de haute dimension). Les soi-disant « hallucinations » des LLM et les possibilités d'injections de prompts ou d'autres attaques

[36] Joshua A. Kroll, « The Fallacy of Inscrutability », *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376.2133 (2018), p. 20180084, doi:10.1098/rsta.2018.0084.

adversaires[37] illustrent assez bien le type de « surprises » qu'une telle opacité peut cacher. Et ces « surprises » que les algorithmes d'apprentissage profond peuvent produire sont très difficiles à prédire et à anticiper *a priori*. Contrairement à la plupart de nos outils (y compris les programmes informatiques conçus de manière plus analytique, comme dans le cas de l'IA symbolique), nous ne disposons pas, dans ce cas, d'une théorie fiable de l'erreur.[38] En ce sens, certains algorithmes d'apprentissage profond peuvent poser un véritable et grave problème de contrôle.

Ce n'est toutefois pas une raison pour simplement interdire ces algorithmes. D'une part, un sous-domaine très actif de l'IA se concentre sur ce sujet de l'explicabilité[39] et des progrès semblent possibles pour interpréter de manière informative ce que nous observons du fonctionnement des programmes de réseaux neuronaux profonds tels que les LLM. D'autre part, rien (sur le plan technique) n'empêche un test empirique *in situ* de ces programmes. Cet aspect est essentiel pour une utilisation éclairée des programmes d'apprentissage profond, en particulier car il permet de distinguer deux grandes configurations. S'il est raisonnable d'espérer rassembler suffisamment d'exemples d'entrées-sorties dans le cas de programmes spécialisés, tels que des outils de classification d'images médicales pour une pathologie spécifique, les choses peuvent se compliquer lorsque la tâche devient plus complexe (classer des images de n'importe quel objet possible) ou même floue (répondre correctement à n'importe quelle question). Plus les résultats attendus varient, plus il deviendra difficile de garantir que les tests empiriques couvrent un champ suffisamment large pour être fiables. Cela signifie que la fiabilité ne peut pas toujours être garantie (en particulier dans le cas de l'IA générative et des LLM) et que les utilisations doivent être adaptées en conséquence. Dans les cas où la fiabilité ne peut être suffisamment garantie, nous devrions nous abstenir de déléguer des tâches sans supervision ainsi que de connecter de manière inappropriée des programmes peu fiables et imprévisibles au reste de nos systèmes d'information (comme cela est parfois promu de manière imprudente avec la tendance croissante de l'IA agentique). C'est typiquement le type de dangers que l'expérience de pensée des « trombones » de Bostrom permet de mettre en é . De tels scénarios sont souvent invoqués pour illustrer la menace que représente une IA devenant plus intelligente que les humains. Non seulement la possibilité, mais aussi la pertinence même de cette idée peuvent être débattues (voir la recommandation 7). Il y a cependant un élément crucial que les scénarios de type « trombones » mettent en évidence : la nature mécanique et inertielle de l'IA.

**Rappel de la matérialité de l'IA**

[37] Yujie Sun et al., « AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content », *Humanities and Social Sciences Communications*, 11.1 (2024), p. 1278, doi:10.1057/s41599-024-03811-x ; Xiaoyong Yuan et al., « Adversarial Examples: Attacks and Defenses for Deep Learning », *IEEE Transactions on Neural Networks and Learning Systems*, 30.9 (2019), pp. 2805–24, doi:10.1109/TNNLS.2018.2886017.

[38] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, NRF Essais (Gallimard, 2023), sec. 4.5, Cairn.info, doi:10.3917/gall.andle.2023.01.

[39] Jacob Dunefsky, Philippe Chlenski et Neel Nanda, « Transcoders Find Interpretable LLM Feature Circuits », *Actes de la 38e Conférence internationale sur les systèmes de traitement de l'information neuronale* (Red Hook, NY, États-Unis), NIPS '24, 37 (2024), pp. 24375–410.

En réalité, l'un des principaux problèmes posés par les scénarios de type « trombones » est que la machine ne peut pas s'écarter de son programme si quelque chose tourne mal. Mais il ne s'agit pas là d'un problème logiciel. Au contraire, une telle inertie est inhérente au matériel et à l'idée même de calcul. Trop souvent, nous imaginons l'IA et les technologies numériques comme largement immatérielles (avec des termes tels que « cloud », « dématérialisation » et des représentations abstraites impliquant des lignes de code colorées, des données ou des équations mathématiques). Cependant, nous ne devons jamais oublier que **tous les programmes (des plus traditionnels et conventionnels aux programmes d'IA les plus avancés issus de l'apprentissage automatique) s'exécutent sur des ordinateurs ou des machines similaires qui ne sont pas (ou peu) programmables**. Pour qu'un tel calcul automatique soit possible, les humains doivent d'abord établir certaines conventions qui associent des significations à des traces ou configurations matérielles (par exemple une série d'aimants sur le disque dur d'un ordinateur dont les orientations symbolisent une séquence de 0 et de 1, elle-même associée, par exemple, à une séquence de mots ou à une séquence de nombres codant les couleurs des pixels d'une image). Ensuite, des humains très compétents peuvent concevoir des machines telles que des ordinateurs qui transformeront de manière précisément contrôlée (en suivant un algorithme ou un programme) ces traces matérielles en de nouvelles traces associées à d'autres significations (par exemple, une nouvelle série de mots, une image modifiée ou une description de l'image). Présenté ainsi, il devient évident que l'une des premières propriétés que l'on attend de telles machines est de ne jamais s'écarter du fonctionnement prévu tel que prescrit par le programme. Ce type de machines, conçues pour transformer des configurations matérielles en d'autres en fonction de ce que ces configurations signifient, n'est pas nouveau. L'ordinateur peut être considéré comme l'aboutissement d'une longue histoire de techniques et de technologies de l'information, remontant probablement aux tout débuts de l'écriture. Dans cette perspective, le boulier peut être vu comme un ancêtre de l'ordinateur (transformation mécanique de configurations symbolisant, par exemple, des nombres à additionner, en configurations symbolisant le résultat de l'addition).

Ainsi, à proprement parler, il n'y a ni significations, ni images, ni mots, ni nombres dans les ordinateurs, et encore moins d'émotions ou de conscience. Ce sont toutefois des machines fantastiques pour manipuler mécaniquement (avec une efficacité et une précision incroyables) d'innombrables configurations matérielles auxquelles nous, les humains, attachons un sens. Une série d'aimants sur le disque dur d'un ordinateur provoquera l'émission de différentes couleurs par différents pixels à l'écran, qui seront pour nous bien plus que de minuscules sources de lumière colorée : ils deviendront des textes nous parlant de sentiments, des images de visages exprimant telles ou telles émotions... Mais l'ordinateur ne fait que traiter des informations en manipulant mécaniquement et automatiquement des aimants (ou d'autres configurations matérielles). Cela rend d'autant plus époustouflant de voir ce que nous pouvons faire faire aux ordinateurs grâce à des programmes dérivés des techniques d'apprentissage automatique.

**Souligner la dépendance directe à l'intelligence humaine**

Il est tout à fait légitime de s'émerveiller devant l'IA (tout comme de s'en inquiéter). Cependant, l'émerveillement que nous pouvons ressentir face à la technologie de l'IA doit être fondé sur

de bonnes raisons. Comme nous venons de le voir, ses succès en matière d' s n'ont rien à voir avec la création de nouvelles formes de vie, de nouveaux êtres intelligents qui nous feraient concurrence, que nous appellerions des IA. Il est tout aussi vertigineux, sinon plus, de réaliser que l'humanité a été capable de construire des machines, des artefacts capables de simuler ou de reproduire un comportement intelligent (un comportement convaincant qui aurait pu provenir d'humains), sans aucune vie, aucune expérience vécue, aucune conscience, mais avec de purs mécanismes (des mécanismes inertes, mais d'une complexité et d'une miniaturisation éblouissantes). Les succès impressionnants de l'IA ne doivent jamais nous faire oublier que l'IA est un produit de l'intelligence humaine, dont elle dépend directement.

L'exposé proposé jusqu'à présent montre clairement, nous l'espérons, que la notion de super-IA (tel que l'a définie Eric Schmidt, un système fournissant des résultats que nous savons vrais sans être capable d'en comprendre la preuve) est largement illusoire. Dans l'état actuel de la technologie de l'IA, il n'existe aucune garantie magique que les résultats produits soient valides. Les techniques d'apprentissage automatique ne changent pas la donne. Les humains restent chargés de créer les programmes informatiques et d'évaluer leurs résultats. Il faut un très haut niveau d'intelligence humaine pour obtenir de bons systèmes d'IA. Bien sûr, cela nécessite des programmeurs, des ingénieurs et des informaticiens très brillants pour concevoir les architectures informatiques et les procédures d'apprentissage. Mais la conception de la rétroaction pour guider le processus d'apprentissage exige également des personnes très brillantes, et ce dans des communautés bien plus larges que le travail de programmation lui-même. La formulation formelle de ce retour d'information peut s'avérer très délicate, nécessitant de nombreuses connaissances de la part d'experts du domaine concerné. La constitution de bases de données fiables d'exemples pour la reconstruction indirecte du retour d'information mobilise l'intelligence de nombreuses personnes (journalistes, contributeurs à Wikipédia, auteurs d'ouvrages, chercheurs publiant des livres ou des articles dans des revues universitaires, ..., toute personne produisant du contenu intelligent sur les plateformes numériques et les systèmes d'information).

La responsabilité humaine irréductible demeure également au niveau de l'évaluation de la qualité et de la fiabilité des systèmes d'IA. Il faut beaucoup d'efforts et d'intelligence humains pour instaurer la confiance dans une technologie numérique donnée. Très souvent, les profanes délèguent ce travail d'évaluation, mais pas aux machines elles-mêmes. Nous le déléguons à des spécialistes en qui nous avons confiance. Et dans le cas de l'IA générative, il est bien établi que le niveau de fiabilité est trop faible pour accorder une confiance aveugle aux résultats (comme l'illustrent très bien les conditions d'utilisation de ces systèmes, qui sont on ne peut plus claires quant à la responsabilité de l'utilisateur vis-à-vis des résultats générés). Pour en revenir au problème du contrôle, il peut être très difficile de garantir que des systèmes génératifs très complexes et puissants ne répondront pas favorablement à des demandes illégitimes ou illégales (telles que le soutien à des actes criminels ou des conseils dangereux à des personnes psychologiquement fragiles). Il est également très difficile de définir de bons objectifs pour les processus d'apprentissage et de fournir de bonnes instructions à des systèmes très complexes et puissants afin de s'assurer qu'ils ne produiront pas de résultats dangereux et inattendus. De plus, il peut être nécessaire de mobiliser beaucoup de ressources humaines et d'intelligence pour s'abstenir d'utiliser, dans des contextes insuffisamment

sécurisés, des systèmes qui n'offriraient pas de garanties suffisantes de fiabilité. D'une manière plus générale, donner une orientation adéquate à l'IA et garantir que les systèmes d'IA respectent nos objectifs présuppose que les humains aient été assez intelligents pour définir ces orientations et ces objectifs (comme évoqué dans les recommandations 1 à 5). C'est peut-être là que le plus haut niveau d'intelligence collective humaine est nécessaire. Comme nous allons le voir maintenant, une entreprise aussi cruciale peut également bénéficier de certaines perspectives anthropologiques et philosophiques.

Ancrer certaines intuitions fondamentales sur les spécificités des humains par rapport aux machines (recommandation 7)

Nous avons développé ce point dans les recommandations 1 à 5, et la section précédente en a encore mieux mis en évidence l'importance. Ce que deviendra l'IA, les services qu'elle nous offrira ou les pièges et menaces qu'elle représentera pour nous sont les conséquences directes de la qualité de la réflexion humaine qui présidera à son développement et à ses utilisations. Cette réflexion, qui vise à affiner notre discernement sur la place de l'IA dans nos sociétés et notre vie quotidienne, nécessite non seulement une compréhension démystifiée de l'IA, mais aussi une approche solide de ce que signifie être humain, de ce qui est au cœur de l'humanité, de nos forces et de nos limites. En particulier, nous avons besoin d'une compréhension solide et partagée de nos spécificités en tant qu'êtres humains par rapport aux systèmes d'IA.

À cet égard, certains experts défendent des thèses très disruptives et contre-intuitives. Dans des interviews récentes, Yan Le Cun ou Stanislas Dehaene ont indiqué qu'ils ne voyaient aucune opposition de principe à ce que les machines deviennent conscientes.[40] Des rapports scientifiques récents ont également affirmé que la possibilité d'une conscience de l'IA ne pouvait être exclue, et qu'il fallait donc considérer sérieusement la question du bien-être de l'IA.[41] Devrions-nous, comme nous y exhorte un nombre croissant de voix, reconsidérer nos intuitions fondamentales sur le statut ontologique des systèmes d'IA et des ordinateurs ? Devrions-nous commencer à envisager que les machines puissent être dotées de traits humains fondamentaux tels que la conscience, l'intentionnalité ou le libre arbitre ? Bien qu'il faille résister au dogmatisme aveugle (ici comme partout ailleurs), la révision d'intuitions et de croyances fondamentales d'une telle importance ne devrait se faire qu'avec une grande prudence, pour de très bonnes raisons. Il est loin d'être évident que ce soit actuellement le cas.

**Résister à l'injonction de l'« outcomisme »**

Bon nombre des arguments qui conduisent à remettre en question les intuitions courantes et à affirmer que l'IA pourrait atteindre un nouveau statut ontologique trouvent leur origine et tirent leur force d'une focalisation exclusive sur les résultats que les humains et les machines

[40] Le Cun (*Le Point*, 2023, https://www.lepoint.fr/sciences-nature/intelligence-artificielle-le-debat-choc-et-inedit-harari-le-cun-11-05-2023-2519779_1924.php) ; Dehaene a déclaré que « la conscience est une propriété computationnelle », *Le Point*, 2023, https://www.lepoint.fr/sciences-nature/intelligence-artificielle-a-quand-une-conscience-artificielle-30-08-2023-2533358_1924.php).

[41] Patrick Butlin et al., « Consciousness in Artificial Intelligence: Insights from the Science of Consciousness », arXiv:2308.08708, prépublication, arXiv, 22 août 2023, doi:10.48550/arXiv.2308.08708 ; Robert Long et al., « Taking AI Welfare Seriously », arXiv:2411.00986, prépublication, arXiv, 4 novembre 2024, doi:10.48550/arXiv.2411.00986.

peuvent produire. En fait, ce raisonnement repose sur l'idée que nous ne pouvons avoir un contact direct avec l'expérience intérieure (conscience phénoménale, volonté, intentionnalité, etc.) que par l'introspection, qui n'est pas une source de connaissance objective et scientifiquement fiable. Comme l'a formulé Chalmers, les aspects phénoménaux de la vie mentale constituent le « problème difficile » ou la « partie difficile » du problème corps-esprit, par opposition aux propriétés psychologiques et aux dimensions comportementales qui peuvent être étudiées objectivement, de l'extérieur, par une approche fonctionnaliste : avec des « propriétés fonctionnelles caractérisées par des rôles causaux, de sorte que la question « Comment un système physique pourrait-il avoir la propriété psychologique P ? » revient à la même chose que « Comment un état d'un système physique pourrait-il jouer tel ou tel rôle causal ? »[42] C'est ce défi lancé à la fiabilité scientifique de l'introspection qui a conduit Turing à proposer le célèbre « jeu de l'imitation » comme seul test légitime pour évaluer si les machines pouvaient penser.[43] Pour le behaviorisme et les formes réductionnistes du fonctionnalisme, les états intérieurs ou mentaux ne peuvent être abordés que comme des liens causaux entre des phénomènes observables (qui peuvent inclure des états cérébraux décrits comme des états internes de machines computationnelles). Comme le dit si bien Janet Levin, les approches de ce type « ne risquent pas de désigner, ni d'induire un engagement envers, des propriétés ou des processus (directement) observables uniquement par l'introspection ».[44]

Ce discrédit épistémique jeté sur l'introspection ouvre la voie à ce que l'on peut appeler l'« outcomisme », qui prescrit de restreindre les discussions sur la possibilité pour les machines de posséder des traits humains spécifiques relevant du domaine phénoménal (conscience, intentionnalité, volonté, …) à la comparaison entre les résultats que les systèmes d'IA et les humains peuvent respectivement produire. Le behaviorisme éthique de Danaher illustre bien cette restriction : nous ne devrions pas demander à l'IA de prouver sa conscience phénoménale d'une manière plus stricte et exigeante que celle que nous appliquons à un ami ou à un autre être humain.[45] Une telle approche « outcomiste » interdit, par principe, toute distinction entre l'émission de mots et de comportements exprimant de la compassion et l'expression d'une expérience compatissante authentique. Cette approche « outcomiste » est très efficace pour ébranler les intuitions les plus profondément ancrées. Les systèmes d'IA générative deviennent de plus en plus capables de passer des tests de Turing artistiques (les humains devenant incapables de déterminer si les productions sont générées par l'IA ou non).[46] Une étude a démontré une tendance à percevoir des intentions et l'expression d'émotions dans des œuvres

---

[42] David John Chalmers, **The Conscious Mind: In Search of a Fundamental Theory**, série Philosophy of Mind (Oxford University Press, 1996), introduction et section I.1.4.

[43] Alan M. Turing, « Computing Machinery and Intelligence », *Mind*, LIX.236 (1950), pp. 433–60, doi:10.1093/mind/LIX.236.433.

[44] Janet Levin, « Functionalism », dans *The Stanford Encyclopedia of Philosophy*, éd. par Edward N. Zalta et Uri Nodelman, été 2023 (Metaphysics Research Lab, Université de Stanford, 2023) <https://plato.stanford.edu/archives/sum2023/entries/functionalism/> [consulté le 28 novembre 2025].

[45] Groupe de recherche sur l'IA du Centre for Digital Culture, « Encountering Artificial Intelligence: Ethical and Anthropological Investigations », *Journal of Moral Theology*, 1.Theological Investigations of AI (2023), pp. 74–80, doi:10.55476/001c.91230.

[46] Brian Porter et Edouard Machery, « La poésie générée par l'IA est indiscernable de la poésie écrite par des humains et est mieux notée », *Scientific Reports*, 14.1 (2024), p. 26133, doi:10.1038/s41598-024-76900-1.

d'art que les gens savent être générées par l'IA.[47] La mentalité « outcomiste » peut même pousser certains auteurs à nier la légitimité de prendre en compte l'identité de l'auteur d'une œuvre d'art lorsqu'on en juge la qualité (ils y voient un préjugé anthropocentrique et affirment qu'il faut « débiaser la perception qu'ont les gens de l'art généré par l'IA », afin de permettre « d'accéder aux choses sous leur forme véritable (…) libérées de tout préjugé et de toute idée préconçue »).[48] Le même raisonnement peut également s'appliquer au domaine moral. Les grands modèles de langage (LLM) semblent capables d'imiter même des experts moraux de renom tels que « The Ethicist » du New York Times.[49] Des cadres fonctionnalistes peuvent être utilisés pour défendre l'attribution d'une capacité d'agir moral à certains systèmes d'IA.[50]

S'il est vrai qu'il existe des débats intenses sur la nature profonde de la conscience phénoménale, de l'intentionnalité et du libre arbitre, ainsi que sur leur attribution éventuelle aux machines, les positions « outcomistes » du type de celles qui viennent d'être évoquées sont loin de faire l'unanimité. Elles ne constituent qu'une partie très limitée de discussions plus larges et très denses.[51] Il ne fait aucun doute que ces débats sont pertinents pour le discernement éthique et ne doivent pas être considérés comme purement philosophiques.[52] Mais il serait trompeur de ne mobiliser qu'une fraction réduite de leurs dimensions argumentatives. Nous devons résister à une telle prise de pouvoir. Encore une fois, il est sain sur le plan cognitif d'être prêt à réviser ses croyances et intuitions fondamentales. Mais nous ne devrions le faire qu'en présence de raisons solides. Il est loin d'être évident que la simple existence et la possibilité de positions philosophiques niant les différences entre les humains et les machines sur ces questions justifient de rejeter les intuitions bien ancrées que nous avons. De plus, les stratégies « outcomistes » discutées ici reposent sur l'hypothèse que la conscience phénoménale, les expériences vécues intérieures et les caractéristiques associées sont des

---

[47] Theresa Rahel Demmer et al., « Does an Emotional Connection to Art Really Require a Human Artist? Emotion and Intentionality Responses to AI- versus Human-Created Art and Impact on Aesthetic Experience », *Computers in Human Behavior*, 148 (2023), p. 107875, doi:10.1016/j.chb.2023.107875.

[48] Kobe Millet et al., « Défendre l'humanité : le biais anthropocentrique dans l'appréciation de l'art généré par l'IA », *Computers in Human Behavior*, 143 (2023), p. 107707, doi:10.1016/j.chb.2023.107707.

[49] Danica Dillion et al., « AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise », *Scientific Reports*, 15.1 (2025), p. 4084, doi:10.1038/s41598-025-86510-0.

[50] Morgan S. Porter, « Moral Agency in Silico: Exploring Free Will in Large Language Models », arXiv:2410.23310, prépublication, arXiv, 28 octobre 2024, doi:10.48550/arXiv.2410.23310.

[51] Voir par exemple : John R. Searle, « Minds, Brains, and Programs », *Behavioral and Brain Sciences*, 3.3 (1980), pp. 417–24, doi:10.1017/S0140525X00005756 ; Hubert Lederer Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (MIT Press, 1992) ; Hubert L. Dreyfus, « Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian », *Artificial Intelligence*, 171.18 (2007), pp. 1137–60, doi:10.1016/j.artint.2007.10.012 ; Adrien Doerig, Aaron Schurger et Michael H. Herzog, « Hard Criteria for Empirical Theories of Consciousness », *Cognitive Neuroscience*, 12.2 (2021), pp. 41–62, doi:10.1080/17588928.2020.1772214 ; Sara Lumbreras et Eduardo Garrido-Merchán, « Insights from Saint Teresa and Saint Augustine on Artificial Intelligence: Discussing Human Interiority », *Scientia et Fides*, 12.2 (2024), pp. 265–95, art. 2, doi:10.12775/SetF.2024.025 ; Andrzej Porębski et Jakub Figura, « Il n'existe pas d'intelligence artificielle consciente », *Humanities and Social Sciences Communications*, 12.1 (2025), p. 1647, doi:10.1057/s41599-025-05868-8.

[52] Mario De Caro, « L'imputabilité nécessite-t-elle le libre arbitre ? Le débat dans la tradition du droit civil », dans *La liberté humaine à l'épreuve de l'IA et des neurosciences*, édité par Stefano Biancu, Mathieu Guillermin et Fabio Macioce, Humanisme contemporain : Annales en libre accès (2024) (Edizioni Studium, 2024), pp. 41–53 <https://www.edizionistudium.it/riviste/studium-contemporary-humanism-open-access-annals-2024>.

trous noirs épistémiques qui ne peuvent être étudiés sérieusement en eux-mêmes. Cette hypothèse même n'est peut-être pas aussi solide qu'elle le semble à première vue.

**Prendre au sérieux l'expérience vécue, la vie et la biologie**

Tout d'abord, on peut souligner le caractère précipité d'un rejet systématique des éléments introspectifs. L'argumentation de Searle dans sa célèbre expérience de pensée de la « chambre chinoise » est particulièrement intéressante dans cette perspective épistémologique.[53] Ce texte est souvent présenté comme une attaque plus ou moins réussie contre le fonctionnalisme. Ce n'est pas cet aspect qui nous intéresse dans le cadre de la présente discussion. Ce qui importe ici, c'est plutôt la place que Searle accorde à l'expérience introspective dans son argumentation, qui peut être reconstituée comme suit : la thèse selon laquelle « avoir un esprit, c'est exécuter le bon type de programme » vise *tous* les esprits (quantification universelle). De là, on peut *déduire* logiquement que la thèse s'applique à tout esprit singulier (le mien, celui du lecteur, celui de Searle). L'expérience de pensée de la chambre chinoise nous permet de nous convaincre par introspection que notre propre esprit ne fonctionne pas ainsi. Chaque personne qui réalise cette expérience introspective prend directement connaissance d'un contre-exemple réfutant la thèse universelle qui s'applique à *tous les esprits*. Compte tenu de la nature de la thèse, une réfutation par l'expérience introspective est parfaitement légitime.

Bien qu'il puisse constituer un élément théorique significatif pour s'opposer à l'outcomisme, cet argument en faveur de la légitimité épistémique de l'introspection ne suffit pas à fournir une base solide pour discuter de la fiabilité de nos intuitions concernant les spécificités que les humains conservent par rapport aux machines en matière de conscience phénoménale, d'intentions ou de libre arbitre. Heureusement, on peut en dire davantage. On peut notamment remettre en question l'affirmation selon laquelle la conscience phénoménale et les caractéristiques mentales associées ne peuvent être étudiées directement de l'extérieur, mais uniquement à travers leurs effets causaux ou réduites à un modèle purement computationnel du cerveau. Une image différente peut émerger si l'on prend en compte les « mécanismes biologiques sous-computationnels ».[54] C'est précisément la piste explorée par Damasio.[55] Pour lui, la conscience phénoménale ne constitue un problème difficile que parce qu'elle est abordée en mettant trop l'accent sur le cerveau et ses propriétés ou descriptions computationnelles. Une fois que l'on accepte d'élargir le champ d'étude, il devient possible de décrire la conscience phénoménale comme ancrée dans l'activité fondamentale des organismes vivants (les processus homéostatiques constituant une sorte d'« intelligence non explicite ») et comme émergeant de l'imbrication riche et organique entre le système nerveux central et le reste du corps vivant. Un élément clé pour Damasio est l'interoception qui, loin d'être une simple représentation mentale de ce qui se passe dans le corps, résulte d'un dialogue dynamique entre les neurones et le reste des tissus et organes vivants. L'interoception constitue la trame

[53] Searle, « Minds, Brains, and Programs ».

[54] Ned Block, « Can Only Meat Machines Be Conscious? », *Trends in Cognitive Sciences*, publié en ligne le 7 octobre 2025, doi:10.1016/j.tics.2025.08.009.

[55] Antonio R. Damasio, *Feeling & Knowing: Making Minds Conscious*, première édition Vintage Books (Vintage Books, 2022).

fondamentale de l'esprit conscient de soi, un arrière-plan de sensations au sein duquel d'autres expériences mentales se produiront.

Malgré leurs limites éventuelles, de telles approches démontrent la possibilité et la légitimité de prendre la vie au sérieux, du moins à travers le prisme de la biologie. Dans cette perspective, les ordinateurs et les machines apparentées appartenant au domaine des technologies de l'information peuvent être clairement distingués des êtres vivants. Ces derniers sont spécifiques du simple fait d'être des organismes vivants. Il ne semble y avoir aucune raison impérieuse d'abandonner nos intuitions fondamentales sur ce qui est vivant ou non (du moins en ce qui concerne les ordinateurs). Dans le même ordre d'idées, nos intuitions communes sur ce qui relève de la conscience et ce qui n'en relève pas semblent suffisamment fiables. Lorsque nous reconnaissons la légitimité de la biologie et des branches des neurosciences qui s'y intéressent encore, les recherches scientifiques peuvent conduire à les affiner, mais en aucun cas à les remettre en cause, en particulier lorsqu'il s'agit d'artefacts issus des technologies de l'information. Le même type de discussion peut être mené sur le thème du libre arbitre. Certes, nous sommes loin de comprendre parfaitement cette notion. Cependant, il semble important de laisser une place aux apports biologiques. Les organismes vivants dotés d'un système nerveux central semblent posséder une sorte d'autonomie et une capacité à contourner les automatismes computationnels inertes dont les automates sont dépourvus.[56]

**Assurer une compréhension solide des spécificités fondamentales de l'être humain**

À la lumière des sections précédentes, il semble important d'approfondir la discussion sur les pratiques linguistiques répandues qui tendent à attribuer à l'IA et aux technologies numériques des traits habituellement associés aux humains (et éventuellement à certains animaux). En effet, il est assez courant de parler de *prise de décision* automatisée ou de LLM révolutionnant le *rapport à la connaissance* ou *à la vérité*. Le terme « intelligence artificielle » lui-même suggère l'idée que nous construisons des machines *intelligentes*. S'il peut être légitime de s'exprimer ainsi selon certains sens ou certaines définitions de ces termes, ce serait une erreur dangereuse de les réduire à ces conceptions limitées. Non seulement cela permettrait d'affirmer que les machines peuvent atteindre (en partie) le statut ontologique des êtres vivants et des humains, mais, ce qui est encore plus inquiétant, cela tendrait à réduire le statut ontologique des êtres vivants et des humains à celui des ordinateurs et des systèmes d'IA. Cela encouragerait à considérer les êtres vivants et les humains comme de simples machines. Néanmoins, nous venons de voir que rien n'empêche de prendre au sérieux la vie et la biologie, ainsi que l'expérience vécue que nous pouvons examiner par l'introspection. Si rien ne l'empêche, c'est peut-être un devoir de le faire. Il semble possible, et donc nécessaire pour le discernement éthique et politique, d'affirmer clairement certaines spécificités fondamentales des êtres humains, au moins par comparaison avec les systèmes d'IA et les ordinateurs.

## Autonomie et prise de décision

[56] Björn Brembs, « Towards a Scientific Concept of Free Will as a Biological Trait: Spontaneous Actions and Decision-Making in Invertebrates », *Proceedings of the Royal Society B: Biological Sciences*, 278.1707 (2010), pp. 930–39, doi:10.1098/rspb.2010.2325.

Dans cette perspective, nous devrions toujours garder à l'esprit que parler de prise de décision automatisée ou d'autonomie de l'IA ne peut être valable que dans un sens très limité. Les ordinateurs et autres artefacts des technologies de l'information sont des machines purement mécaniques et inertes. Ainsi, ce que nous entendons par autonomie des machines ne peut être rien de plus qu'une réaction plus ou moins complexe aux variations des données qu'elles reçoivent. Bien sûr, certains programmes peuvent mettre à jour la mémoire interne d'une machine informatique qui peut alors réagir différemment à de nouvelles données. Certains programmes peuvent même modifier d'autres programmes ou sélectionner quel programme exécuter dans des circonstances données. Mais tout cela, considéré globalement, reste des réponses déterministes et mécanistes. Un ordinateur dans un état interne donné réagit toujours de la même manière à la même entrée (du moins, c'est ce que nous attendons des ordinateurs, la raison pour laquelle nous les construisons… et nous déployons beaucoup d'efforts pour compenser lorsqu'ils fonctionnent mal à cet égard, par exemple avec des codes de correction d'erreurs). En termes d'autonomie, les ordinateurs, avec ou sans programmes d'IA, appartiennent au même domaine que les thermostats. Ils ne diffèrent que par le degré de complexité de leurs réponses.

Lorsque nous parlons d'autonomie et de prise de décision chez les êtres vivants et les humains, nous entendons quelque chose de plus que cela, quelque chose de plus fort, d'ontologiquement différent. En particulier lorsqu'il s'agit des humains (dont chacun de nous connaît l'autonomie et le processus décisionnel de l'intérieur grâce à l'introspection), il semble assez évident qu'être autonome ou décider ne se résume pas à appliquer des algorithmes ou des procédures de manière mécanique. Dotés de leur type d'autonomie propre, les êtres vivants peuvent réagir différemment à des sollicitations identiques. Les êtres vivants conscients, comme les humains, peuvent le faire grâce à leur autonomie pratique et à leur libre arbitre. C'est du moins ce que signifie la prise de décision au sens fort (un sens que, comme nous venons de le voir, rien ne nous oblige à abandonner ou à refuser aux humains) : *choisir* volontairement parmi les options disponibles. Les humains ont la capacité de décider de ce qui *devrait être* lorsqu'ils sont confrontés à une pluralité de possibilités. Et comme Damasio l'a démontré de manière célèbre avec l'hypothèse des « marqueurs somatiques », cette capacité à décider au sens fort ne repose pas sur des caractéristiques computationnelles que notre cerveau pourrait posséder, mais implique de manière irréductible des affects et des émotions.[57]

Ce que nous venons d'exposer s'applique également à la prise de décision morale. Produire (comme les LLM semblent en être capables) des phrases qui véhiculent des conseils ou des jugements moraux n'est pas la même chose que posséder véritablement une expertise morale et être capable de prendre des décisions morales au sens fort. Et la question de déterminer si le contenu moral des phrases produites par les LLM est valide ou non passe à côté de l'essentiel. La justesse de ce contenu est pertinente pour un autre sujet (voir ci-dessous la discussion sur le rôle des conseillers moraux artificiels), mais elle est orthogonale à la présente réflexion. Lorsque les humains prennent des décisions dans le domaine moral (comme dans tout autre domaine), ils font plus que simplement réappliquer des réponses passées à des problèmes moraux. Ils mobilisent leur capacité à prendre du recul, à véritablement examiner

[57] Antonio R. Damasio, *L'erreur de Descartes : Émotion, raison et cerveau humain* (Penguin, 2005).

les options. Ils font de leur mieux pour choisir la meilleure. Ils peuvent également sentir que quelque chose ne va pas dans les anciennes façons de faire ou les anciennes normes. Comme le dit Dominique Lambert, les humains sont capables de « créativité » au sens d'un « pouvoir de créer de la nouveauté ».[58] Ce pouvoir de contourner, de prendre de la distance par rapport aux régularités du passé est essentiel d'un point de vue moral. Les techniques d'apprentissage automatique peuvent permettre aux systèmes d'IA de prédire avec justesse ce que les gens pourraient faire en se basant sur ce qu'ils ont fait dans le passé. Mais, comme l'a rappelé avec force le pape François, « on ne doit pas permettre aux algorithmes (…) d'éliminer la possibilité pour un individu de changer et de laisser son passé derrière lui ».[59] Seuls les humains sont capables de rester ouverts et de faire face à de telles possibilités. Un tel pouvoir créatif est au cœur de la prise de décision (morale) au sens fort.

### Relation avec la connaissance et la vérité

Nous pouvons maintenant aborder le sujet de la connaissance et de la relation à la vérité. Comment comprendre les affirmations selon lesquelles l'IA révolutionne notre accès à la connaissance ? Est-il pertinent de penser que l'IA pourrait produire une meilleure connaissance que nous ? Devrions-nous vraiment revoir nos intuitions fondamentales sur la connaissance et la relation à la vérité en tant que traits spécifiques essentiels des êtres humains ? Nous avons déjà fourni quelques éléments qui tempèrent cette question en nous appuyant sur la réalité des techniques d'apprentissage automatique (voir la recommandation 6). Mais on peut en dire davantage en examinant ce que signifie, au sens fort, pour les humains de savoir et d'avoir une relation à la vérité. Ici, comme précédemment, il est crucial de résister à toute approche « outcomiste ». La connaissance n'est pas simplement un ensemble d'énoncés vrais. Produire de la connaissance ne se réduit pas à élaborer des affirmations vraies. Les définitions traditionnelles sont claires à ce sujet : « la connaissance est une croyance vraie *justifiée* ».[60] De même, « une personne, *S*, sait que *p* (où *p* est une proposition) si et seulement si (i) *S* croit que *p*, (ii) *S* a une justification (preuve, bonnes raisons) pour *p*, et (iii) *p* est vrai ».[61] La connaissance nécessite une justification, de bonnes raisons de croire une affirmation donnée.

Comme le rappelle la recommandation 6, on ne peut pas faire aveuglément confiance à la technologie actuelle des LLM. On ne peut pas dire que les LLM produisent elles-mêmes de la connaissance, car les humains doivent toujours vérifier eux-mêmes la validité des résultats. Cependant, les prophètes de l'IA superintelligente ne prétendent pas que la superintelligence est déjà là. Ils affirment simplement que c'est possible et que cela se produira dans un avenir prévisible. Il se peut que ces prédictions tirent leur force de l'idée que les ordinateurs,

---

[58] Notre traduction. Dominique Lambert, « Retrouver l'humain au cœur de l'IA et de la robotique, et lui redonner toute sa place : Conférence », *Revue Confluence*, n° 6.2 (2024), p. 23-42, doi:10.3917/confl.006.0023.

[59] Pape François, *Intelligence artificielle et paix*, Message pour la[57e]Journée mondiale de la paix (1er janvier 2024), https://www.vatican.va/content/francesco/en/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html

[60] James Ladyman, *Understanding Philosophy of Science* (Routledge, 2002), p. 6.

[61] Gürol Irzik, « Critical Rationalism », dans *The Routledge Companion to Philosophy of Science*, édité par Martin Curd et Stathis Psillos, Routledge Philosophy Companions, deuxième édition (Routledge, 2014), pp. 70–78.

précisément parce qu'ils ne sont pas vivants, disposent d'une longueur d'avance de principe sur les humains. Comme ils n'appliquent que des opérations logico-mathématiques à des données brutes, les machines de calcul seraient dotées d'une sorte d'objectivité parfaite, d'une forme supérieure de rationalité libérée par principe de toute arbitraire ou subjectivité. À ce stade, il est important de préciser qu'un tel raisonnement repose sur une conception très spécifique (bien que répandue) de l'intelligence et de la rationalité : être rationnel ou objectif, c'est purger les procédures d'investigation de tout contenu contingent, de tout élément qui pourrait être différent et qui nécessiterait de faire un choix et donc d'évaluer des options. Dans une telle conception de la rationalité comme « pure enquête »,[62] toute spécificité des sujets doit être éliminée. La connaissance et la vérité doivent être recherchées à travers une sorte d'« objectivité mécanique » fondée exclusivement sur des mesures empiriques et des procédures algorithmiques ou logico-mathématiques.[63] En s'appuyant sur ce type d'approche, il est en effet tentant d'imaginer que des machines dotées de puissances de calcul et de volumes de données toujours croissants pourraient, à un moment donné, atteindre une forme supérieure de rapport à la connaissance et à la vérité.

Cependant, l'histoire et la philosophie récentes de la science (depuis au moins la seconde moitié du XXesiècle) nous ont montré les limites d'une telle conception purement algorithmique ou procédurale de la rationalité et de l'intelligence. La notion de justification (de bonne raison de considérer une croyance comme un savoir) ne se prête pas entièrement à une objectivité mécanique. Tout effort visant à élaborer une connaissance, même les plus scientifiques et expérimentaux, englobe inévitablement un espace irréductible de liberté et implique une activité inéliminable de jugement informel de la part des sujets connaissants. Il n'existe pas de données brutes ni de procédures logico-mathématiques neutres qui s'imposeraient d'elles-mêmes. Les jugements et arbitrages humains sont indispensables (par exemple concernant le vocabulaire de base à utiliser, les grandes orientations méthodologiques, les objectifs à atteindre… mais aussi concernant des intuitions fondamentales telles que l'idée que l'observation empirique ne nous trompe pas systématiquement).[64] Comme le dit Hilary

[62] Bernard Arthur Owen Williams, *Descartes : The Project of Pure Enquiry*, [éd. rév.] (Routledge, 2005).

[63] Reiss et Sprenger, « Scientific Objectivity », sec. 4.

[64] Kitcher, *Science, Truth and Democracy*, p. Voir par exemple :; Ernan McMullin, « The Virtues of a Good Theory », dans *The Routledge Companion to Philosophy of Science*, sous la direction de Martin Curd et Stathis Psyllos, Routledge Philosophy Companions, 2e éd. (Routledge, 2014), p. 561–72; Mathieu Guillermin, « Non-neutralité sans relativisme ? Le rôle de la rationalité évaluative », dans *Et si la recherche scientifique ne pouvait pas être neutre ?*, sous la direction de Laurence Brière, Mélissa Lieutenant-Gosselin et Florence Piron (Éditions Science et bien commun, 2019), pp. 315–38 ; Pierre-Luc Dostie Proulx et Mathieu Guillermin, « The Role of Explanatory Virtues in Abduction and IBE », dans *Logic in Question*, sous la direction de Jean-Yves Béziau et al., Studies in Universal Logic (Springer International Publishing, 2022), pp. 471–90, doi:10.1007/978-3-030-94452-0_24 ; Kyle Stanford, « Underdetermination of Scientific Theory », dans *The Stanford Encyclopedia of Philosophy*, éd. par Edward N. Zalta et Uri Nodelman, été 2023 (Metaphysics Research Lab, Université de Stanford, 2023) <https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination/> [consulté le 2 juin 2025] ; Catherine Allamel-Raffin, « Modes de raisonnement et images scientifiques », *Visible*, n° 11 (mai 2023), doi:10.25965/visible.165.

Putnam, avec Cavell, la connaissance, et plus largement « les sujets parlants et pensants » s, ne peuvent exister sans des actes contextualisés et situés de « reconnaissance » et d'« accord ».[65]

Par conséquent, être intelligent ou rationnel, entretenir une relation avec la connaissance et la vérité, c'est bien sûr être capable d'appliquer correctement (objectivement ou neutrement) des critères, des procédures ou des algorithmes, mais c'est aussi, et peut-être surtout, être capable de juger de la qualité des critères et des procédures, d'avoir une attitude réflexive et critique envers ce que nous faisons… et donc être capable de juger et d'arbitrer de manière faillible, de se tromper parfois, de se corriger, d'évoluer (et de s'entraider à cet égard, de collaborer de bonne volonté)… Être intelligent dans ce sens fort est quelque chose de fondamentalement vivant, quelque chose que chacun de nous ne peut entreprendre qu'en s'enracinant dans sa propre expérience vécue (avec toute la richesse mais aussi les limites que cela implique)[66] et dans une collaboration saine avec les autres. De plus, ce que nous venons d'exposer met en lumière le lien intime entre le fait d'entretenir une relation avec la connaissance et la vérité au sens fort et la capacité de prendre des décisions (également au sens fort). Savoir, au sens humain, implique de manière irréductible d'être confronté à des options disponibles parmi lesquelles il faut choisir. Cela implique une autonomie pratique au sein d'un espace de liberté essentiel. Les humains peuvent avoir une relation à la connaissance et à la vérité grâce à leur capacité à prendre du recul, à concevoir que les choses (ici leurs représentations et leurs croyances admises) pourraient être différentes de ce qu'elles sont. Seule cette capacité place les humains en position de savoir au sens fort, ce qui implique de juger au mieux, sans certitude absolue, laquelle des options disponibles semble la plus raisonnable. Seule cette capacité à prendre du recul rend les humains sensibles à l'appel à faire un usage responsable de leur liberté et de leur autonomie pratique dans une quête sincère de la vérité.

## Quelques thèmes clés à explorer collectivement

À la lumière de ce qui a été discuté jusqu'à présent, nous pouvons affiner notre compréhension du défi que représente la réglementation de l'IA. Il serait dangereux de réduire ce défi à une question d'asymétries de pouvoir entre les pays, entre les géants de la technologie et les utilisateurs, etc., comme si l'orientation que nous devrions donner à l'IA était évidente et que le problème consistait uniquement à neutraliser des acteurs malveillants qui ont de forts intérêts à pousser dans d'autres directions. L'hypothèse selon laquelle nous savons où l'IA devrait aller est plus que discutable. Une part considérable du défi de la réglementation de l'IA consiste en effet à définir les buts et objectifs que l'IA devrait servir. Certains pourraient tenter de propager l'idée selon laquelle l'IA (générale ou super-IA) est une fin en soi, car, une fois atteinte, elle aurait le potentiel de résoudre tous nos problèmes. Cependant, il devrait désormais être suffisamment clair que ce type d'IA magique relève purement de la fantaisie, compte tenu de l'état actuel et prévisible de la technologie de l'IA. Il ne fait aucun doute que

[65] Hilary Putnam, « Philosophy as the Education of Grownups : Stanley Cavell and Skepticism », dans *Philosophy in an Age of Science: Physics, Mathematics, and Skepticism*, éd. par David MacArthur et Mario De Caro (Harvard University Press, 2012), pp. 552–64.

[66] François Laplantine, *La vie des sens : introduction à une anthropologie modale*, Sensory Studies Series (Bloomsbury Academic, 2015).

les techniques d'apprentissage automatique et les systèmes d'IA peuvent résoudre de nombreux problèmes et contribuer à atténuer les enjeux civilisationnels les plus aigus auxquels nous sommes confrontés. C'est déjà le cas. Mais l' e de l'IA ne peut être utile en ce sens que si nous sommes capables de définir ce que nous attendons d'elle, d'affiner nos objectifs et la manière dont nous voulons les atteindre. Et cela est loin d'être toujours simple.

Les problèmes auxquels nous sommes confrontés sont souvent très complexes et, parfois, une réponse trop hâtive faisant appel à l'IA peut s'avérer inadaptée (voire aggraver les problèmes existants). Prenons l'exemple du manque de personnel dans les maisons de retraite. Bien sûr, certains robots bien conçus peuvent aider le personnel soignant à gagner du temps et à travailler plus rapidement. Mais ne devrions-nous pas également réfléchir à la cause de ce manque de personnel ? Ce travail est-il suffisamment valorisé ? Suffisamment rémunéré ? Le développement et le déploiement de solutions d'IA devraient toujours s'inscrire dans le cadre de processus de conception approfondis impliquant les parties prenantes concernées, afin d'affiner la compréhension du problème spécifique à traiter et de déterminer ce qui pourrait constituer une véritable solution ou un moyen de l'atténuer. Ces processus devraient également accorder une attention particulière au choix de la technologie appropriée et à son adaptation à l'objectif spécifique visé. Cela signifie notamment ne pas chercher à mettre en œuvre sans discernement des outils d'IA générative ou d'apprentissage profond. Ceux-ci peuvent parfois s'avérer inappropriés. Parfois, les utiliser à la place de programmes plus simples revient à utiliser un marteau pour casser une noix.

De tels processus de conception sont essentiels pour créer une véritable valeur avec l'IA. Ils permettent de faire évoluer la question générale de « quelle est notre place en tant qu'êtres humains dans le nouveau monde façonné par l'IA ? » vers « quelle est la place de l'IA dans le monde que nous voulons construire, dans notre monde humain, parmi les autres êtres vivants ? ». Dans les contextes locaux, cela implique que les communautés concernées s'engagent dans une réflexion collective approfondie sur le problème auquel elles sont confrontées et sur la manière dont elles attendent de l'IA qu'elle aide à le résoudre. C'est là que réside le cœur de l'exploration éthique et politique que les communautés sociétales doivent mener pour apporter des contributions fructueuses et une réglementation au développement et à l'utilisation de l'IA. Ces réflexions abordent de nombreux thèmes clés différents. Nous ne développerons pas ici ceux qui ont déjà été largement abordés (tels que les inégalités, les questions environnementales, les questions relatives à la protection de la vie privée, la question du modèle économique et de la propriété intellectuelle, ...).[67] Nous souhaitons plutôt mettre en lumière l'importance d'une réflexion de fond sur ce que signifie être humain. Comme nous

[67] Voir par exemple : OCDE, *Mesurer les impacts environnementaux des calculs et des applications de l'intelligence artificielle : l'empreinte de l'IA*, Documents de l'OCDE sur l'économie numérique n° 341, Documents de l'OCDE sur l'économie numérique (2022), CCCXLI, doi:10.1787/7babf571-en ; Organisation internationale du travail et Nations Unies, *Attention à la fracture de l'IA : définir une perspective mondiale sur l'avenir du travail* (2024) <https://www.ilo.org/publications/major-publications/mind-ai-divide-shaping-global-perspective-future-work> [consulté le 1er décembre 2025] ; OCDE, *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions* (2025), doi:10.1787/795de142-en ; UNESCO, *AI and Education: Protecting the Rights of Learners* (2025), doi:10.54675/ROQH4287 ; AIE, *Énergie et IA – Analyse* (2025) <https://www.iea.org/reports/energy-and-ai> [consulté le 1er décembre 2025].

l'avons vu, les machines et les humains ne sont pas interchangeables à bien des égards. Dans certains cas (comme en matière de prise de décision), déléguer à des systèmes d'IA revient à écarter radicalement l'action humaine pour la remplacer par quelque chose de loin d'être équivalent. Cela soulève le défi particulièrement aigu de déterminer comment positionner l'IA pour qu'elle préserve, voire favorise, l'épanouissement des spécificités fondamentales de l'humain.

Explorer les moyens d'aider et de soutenir les êtres humains dans leur rapport à la connaissance et à la vérité (recommandation n° 8)

Comme nous l'avons vu, l'une des caractéristiques fondamentales des êtres humains est leur capacité, bien que faillible, à se rapporter à la vérité, en utilisant leur intelligence pour construire des corpus de croyances qu'ils jugent dignes du titre de savoir. Mettre l'IA véritablement au service de l'intelligence humaine est loin d'être simple et exigera un effort de discernement approfondi pour déterminer où et comment l'IA peut contribuer positivement (ou, au contraire, nuire) aux efforts humains visant à élaborer le savoir et à se rapporter à la vérité.

Une première initiative essentielle consiste à développer un sens individuel et collectif solide permettant de discerner si et comment un système d'IA donné, dans un contexte spécifique, peut être une source de connaissance. Les clarifications et rappels fournis dans les recommandations 6 et 7 sont cruciaux à cet égard. Il n'existe aucune garantie magique que les systèmes d'IA produiront des résultats valides méritant d'être considérés comme des éléments de connaissance. L'IA ne peut se justifier d'elle-même. Seuls les humains possèdent l'autonomie et la capacité de juger si les raisons disponibles en faveur d'une production (humaine ou issue de l'IA) sont suffisantes. Dans ces domaines, la division du travail et la délégation sont courantes et indispensables. Dans de nombreux domaines des technologies de l'information, nous (les utilisateurs finaux) faisons confiance à des groupes d'experts pour évaluer les outils que nous utilisons et la fiabilité de leurs résultats. Bien que nous puissions constater par nous-mêmes quand notre ordinateur ou Internet tombe en panne, nous déléguons largement cette tâche et faisons confiance aux entreprises informatiques et de télécommunications pour fournir des appareils fiables. De même, nous faisons confiance aux éditeurs de logiciels pour fournir des traitements de texte ou des tableurs efficaces, entre autres. En fonction des contenus auxquels nous accédons grâce aux technologies numériques, nous faisons généralement confiance aux fournisseurs de contenu pour partager des informations vérifiées (revues en ligne, encyclopédies, etc.). Nous attendons de toutes ces personnes et de tous ces groupes en qui nous avons confiance qu'ils s'acquittent de leur devoir, afin de garantir que les systèmes auxquels ils participent produisent des connaissances authentiques.

Il est intéressant de noter à cet égard que l'IA générative constitue une exception. Rappelons-le encore une fois (mieux vaut prévenir que guérir) : la validité des résultats des LLM n'est pas garantie. Il n'existe aucun sous-groupe spécifique d'humains chargé (ayant le devoir) de vérifier le contenu particulier qui est fourni à un utilisateur final donné (seuls certains échantillons sont testés, notamment pendant la phase d'entraînement). Cela fait une différence décisive par rapport à ce à quoi nous nous attendons et ce que nous obtenons habituellement, du moins

de la part des fournisseurs et des contenus auxquels nous faisons confiance (par exemple une encyclopédie fiable). Dans ces cas-là, quelqu'un a vécu une expérience lors de l'élaboration et/ou de l'évaluation d'un contenu donné, une expérience cognitive à travers laquelle il ou elle a tenté de produire un savoir authentique, en entretenant une relation authentique avec la vérité. C'est ce type d'expérience auquel nous faisons confiance lorsque nous déléguons l'évaluation de la fiabilité à d'autres êtres humains. Vu sous cet angle, ce que les humains ont accompli en termes d'élaboration et de partage des connaissances, grâce à leur coopération et avec le soutien des technologies de l'information, est absolument stupéfiant. Il ne s'agit pas seulement d'une multitude de contenus différents, sur une multitude de sujets différents, accessibles à presque tout le monde. Mais d'une multitude d'expériences vécues en profondeur, d'élaboration et de validation des connaissances, qui sont mises en commun. C'est là le point clé merveilleux. Et c'est un élément que l'IA ne peut pas reproduire. Au lieu d'un réseau solide, partagé et collaboratif de travail cognitif authentique et d'expériences de rapport à la vérité qui permettent d'établir une confiance légitime, les LLM fournissent aux utilisateurs finaux des résultats qui sont souvent vrais mais qui ne sont jamais garantis (par un humain). La charge d'évaluer ces résultats est transférée, qu'ils en soient conscients ou non, aux utilisateurs finaux.

Bien sûr, la technologie de l'IA ne se réduit pas aux LLM et à l'IA générative. De nombreux outils numériques et d'IA sont jugés fiables et utilisés dans une grande variété de contextes, même (ou surtout) dans les domaines scientifiques. Mais ce que nous avons dit permet de souligner l'importance de la culture de l'IA. Les utilisateurs finaux doivent être conscients des forces et des limites des systèmes qu'ils utilisent, ainsi que des rôles que ces systèmes peuvent jouer ou non dans les processus de construction des connaissances. Les utilisateurs finaux ont un rôle à jouer pour préserver et prolonger les efforts visant à se rapprocher de la vérité. Pour ce faire, ils doivent être capables de discerner s'ils peuvent considérer les résultats produits comme des éléments de connaissance (parce qu'ils font confiance à d'autres humains qui les garantissent) ou si un travail d'évaluation reste à faire. Il serait très dangereux de présupposer que cette capacité est déjà suffisamment développée. Au contraire, des études empiriques suggèrent le contraire (les profanes ont tendance à surestimer la fiabilité des LLM).[68] Il est donc très important de favoriser la culture de l'IA en lien avec ces enjeux épistémiques. Et il ne s'agit pas ici de dire que les systèmes auxquels nous ne pouvons pas faire confiance en tant que sources fiables de connaissances devraient être interdits par principe. Il peut être très judicieux de mobiliser l'IA générative pour nous aider dans nos tâches exploratoires, même lorsque nous savons que nous ne pouvons pas nous fier aveuglément à ses résultats, tant que nous nous abstenons de nous y fier entièrement lorsqu'il s'agit de justifier qu'une chose mérite d'être considérée comme un véritable élément de connaissance.[69] Notre argument est plutôt que nous devons développer notre capacité à attribuer à la technologie de l'IA la place qui lui revient dans nos processus d'élaboration des connaissances et dans la manière dont nous abordons notre rapport à la vérité. Cela implique d'avoir une vision lucide de l'état actuel des

[68] Steyvers et al., « Ce que savent les grands modèles linguistiques et ce que les gens pensent qu'ils savent » ; Union (UER), *L'intégrité de l'information dans les assistants IA*.

[69] C'est l'idée qui sous-tend la distinction, classique en épistémologie, entre les contextes de découverte et de justification. Voir par exemple : Ladyman, *Understanding Philosophy of Science*, sec. 3.3.

technologies d'IA, mais aussi d'explorer le type de nouveaux systèmes que nous pourrions développer pour apporter un soutien supplémentaire dédié aux diverses facettes de nos vies cognitives et épistémiques.

Réfléchir à la manière dont l'IA peut servir l'intelligence humaine (collective) (recommandation 9)

Il est essentiel de renforcer notre capacité à évaluer correctement la légitimité de considérer certains systèmes d'IA comme des sources de connaissance. Cependant, cela ne doit pas épuiser la réflexion sur le discernement. Au-delà de la question de la fiabilité des outils, se pose une question plus globale concernant la contribution de l'IA à la préservation et au développement de l'intelligence humaine (individuelle et collective). Il s'agit à la fois d'un des objectifs fondamentaux que l'IA devrait servir (celui sur lequel nous nous concentrons dans cette section) et d'une condition préalable au développement de programmes utiles et à leur utilisation appropriée (comme nous venons de le voir). Dans cette perspective plus globale, la fiabilité d'un système d'IA en termes de production ou de transmission de connaissances ne garantit pas de manière absolue qu'il contribuera à l'épanouissement de la vie intellectuelle et cognitive des êtres humains. En effet, un nombre croissant de données suggère que la surutilisation de systèmes d'IA générative tels que les LLM peut conduire à une déqualification ou entraver le développement cognitif.[70] S'il peut être tout à fait acceptable, voire souhaitable, de cesser d'accomplir certaines tâches, la délégation à l'IA (générative) doit toujours s'accompagner d'une analyse approfondie des compétences que nous risquons de perdre ou de ne pas développer, ainsi que d'un discernement quant à l' e caractère dispensable de ces compétences (non seulement pour continuer à pouvoir vérifier les résultats de l'IA, mais aussi parce que certaines tâches et compétences peuvent être impliquées dans d'autres domaines du développement cognitif des personnes).

La contribution de l'IA à la préservation et au développement de l'intelligence humaine doit également être abordée de manière plus globale, en s'éloignant de l'IA générative et des modèles de langage (LLM). Comme déjà évoqué, les outils numériques et d'IA ne peuvent être considérés comme de simples instruments que nous pourrions décider d'intégrer ou non dans nos processus intellectuels et cognitifs. Cela est certes vrai dans certains cas, mais ne peut être généralisé. En effet, la technologie numérique est également devenue un milieu dans lequel nous menons notre vie cognitive et nos efforts d'élaboration des connaissances. S'appuyant sur de puissants algorithmes prédictifs, les moteurs de recommandation et de recherche éditorialisent pour nous la vaste quantité d'informations disponibles sur Internet. De plus en plus, l'IA générative est mobilisée pour organiser et résumer des contenus trop lourds pour être traités directement. Les systèmes d'IA deviennent ainsi une sorte d'« extension cognitive » de l'esprit humain. Ils « façonnent de plus en plus le substrat informationnel sur lequel opère

[70] Muhammad Abbas, Farooq Ahmed Jam et Tariq Iqbal Khan, « Is It Harmful or Helpful? Examining the Causes and Consequences of Generative AI Usage among University Students », *International Journal of Educational Technology in Higher Education*, 21.1 (2024), p. 10, doi:10.1186/s41239-024-00444-7 ; Nataliya Kosmyna et al., « Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task », arXiv:2506.08872, version 1, prépublication, arXiv, 10 juin 2025, doi:10.48550/arXiv.2506.08872.

la cognition humaine ».[71] Ici, l'action des systèmes d'IA passe, pour une grande part, inaperçue. Leurs résultats dans ce domaine constitueront donc une condition préalable, un terrain plus ou moins fertile pour l'intelligence humaine.

Comme nous l'avons rappelé dans la recommandation n° 7, les processus que les êtres humains mettent en œuvre pour construire le savoir, ainsi que leur rapport à la vérité, sont loin d'être infaillibles. Les êtres humains doivent s'efforcer de faire un usage responsable de leur liberté de pensée, afin de juger, à partir de leur expérience vécue et incarnée, si une croyance donnée s'accompagne de raisons suffisamment solides pour être considérée comme une connaissance. Une partie importante de cette relation vécue à la vérité et à la connaissance est collective. Savoir, ce n'est pas seulement se convaincre, de manière isolée, que quelque chose est valable. C'est aussi s'engager sur la validité de ce que nous considérons comme une connaissance devant les autres. Lorsque nous pensons savoir quelque chose, nous attendons des autres qu'ils soient d'accord avec nous. Le fait qu'ils soient d'accord avec nous a de l'importance. S'ils ne le sont pas, c'est une bonne raison de douter. C'est là le cœur de l'élaboration de la connaissance scientifique (avec des communautés scientifiques organisées pour favoriser de tels processus collectifs), mais cela concerne aussi (d'abord et peut-être avant tout) les jugements et les engagements à l'égard des preuves fondamentales et factuelles relevant du bon sens (comment les communautés scientifiques pourraient-elles fonctionner correctement sans ce fondement fondamental). Une grande partie de la qualité de la vie intellectuelle et cognitive des êtres humains repose sur cette « décence commune », cette volonté fondamentale des êtres humains de répondre à l'appel du jugement et de la connaissance en commun.[72]

À cet égard, nous devons donc nous demander si les contributions des systèmes d'IA, en tant qu'extension cognitive ou milieu de conditionnement, sont positives ou non. Nous fournit-on les informations et les connaissances les plus utiles possibles ? Les systèmes d'IA favorisent-ils la compréhension mutuelle et l'enrichissement ? Sont-ils au service d'un véritable effort collectif humain visant à se rapprocher de la vérité et à produire des connaissances ? Bien que le potentiel positif de l'IA soit indéniable et particulièrement riche, certaines contributions déjà effectives des systèmes d'IA à la structuration de nos vies cognitives collectives sont extrêmement préoccupantes. Comme on le sait bien, l'éditorialisation automatisée de nos paysages informationnels conduit souvent à des bulles cognitives, des chambres d'écho où la polarisation de l' s se déchaîne et défait le réseau épistémique humain. Cela a conduit certains chercheurs à proposer les concepts de préjudice épistémique et d'injustice épistémique, consistant en la dégradation illégitime, due aux technologies numériques, du « statut épistémique » des personnes (la perception que l'on a de soi-même et que les autres ont de notre capacité à connaître, à interpréter et à témoigner fidèlement).[73] Dans de tels contextes

[71] Massimo Chiriatti et al., « System 0 : Transforming Artificial Intelligence into a Cognitive Extension », *Cyberpsychology, Behavior, and Social Networking*, vol. 28, n° 7 (2025), p. 534-542, doi:10.1089/cyber.2025.0201.

[72] Revault d'Allonnes, *La faiblesse du vrai*.

[73] John Symons et Ramón Alvarado, « Epistemic Injustice and Data Science Technologies », *Synthese*, 200.2 (2022), p. 87, doi:10.1007/s11229-022-03631-z ; Jackie Kay, Atoosa Kasirzadeh et Shakir Mohamed, « Epistemic Injustice in Generative AI », dans *Actes de la conférence AAAI/ACM 2024 sur l'IA, l'éthique et la société* (AAAI Press, 2025), pp. 684–97.

post-vérité, la confiance et la bienveillance nécessaires pour préserver et favoriser une vie cognitive collective fructueuse sont écartées et sapées.

Ces contributions toxiques des systèmes d'IA ne sont pas une fatalité. Elles ne sont pas intrinsèques à l'IA employée pour organiser notre milieu cognitif. L'IA possède un potentiel énorme pour favoriser l'intelligence collective. Elle pourrait nous recommander des informations personnalisées dont nous avons besoin et qui peuvent élargir nos perspectives, nous rendant plus réceptifs à des rencontres fructueuses avec des opinions et des idées divergentes. Elle pourrait nous aider à consacrer plus de temps à des relations authentiques avec les autres (notamment en nous aidant à nous éloigner de nos écrans de temps à autre). Cependant, ces perspectives prometteuses se heurtent à la réalité du modèle économique « gratuit » qui domine largement le secteur des technologies numériques (du moins en matière de logiciels). En effet, ce modèle économique repose sur la captation de l'attention des individus, un objectif qui préside à la conception des systèmes d'IA conditionnant notre environnement cognitif et qui est en grande partie responsable des problèmes aigus que nous venons d'évoquer.[74] Des simulations récentes suggèrent qu'il ne s'agit pas de quelque chose qui peut être réglementé de l'extérieur, les problèmes étant intrinsèquement liés à l'objectif fondamental de la captation de l'attention.[75] Cela signifie que pour tirer le meilleur parti de ce que l'IA peut apporter en tant qu'extension cognitive, nous aurons besoin de nombreux efforts collectifs et d'une grande intelligence pour réorienter nos modèles économiques et nos pratiques de consommation. Là encore, nous voyons la dimension politique profonde de l'IA et la responsabilité collective qui en découle.

Explorer comment l'IA peut contribuer à l'action et à la responsabilité humaines (recommandation 10)

Avec la complexité croissante des technologies numériques (et des systèmes socio-techniques associés), un débat intense a émergé, soulignant les problèmes potentiels liés à l'attribution de la responsabilité (morale ou juridique) en cas de préjudice ou de problème généré par des systèmes avancés (en particulier ceux basés sur des techniques d'apprentissage automatique). Une grande partie du débat repose sur la reconnaissance commune que l'IA et les technologies numériques ne peuvent être considérées comme décidant ou agissant au sens strict impliquant responsabilité et obligation de rendre compte (comme nous le rappelons dans les recommandations 6 et 7). Ainsi, si l'IA ne peut être responsable alors qu'elle complexifie et obscurcit néanmoins les schémas causaux, des lacunes en matière de responsabilité peuvent apparaître, empêchant l'attribution de la responsabilité.[76]

Bien que ces difficultés potentielles d'attribution de la responsabilité soient très importantes (en particulier d'un point de vue juridique), nous souhaitons ici les aborder de manière indirecte, plus en phase avec le cœur de notre discussion. En effet, certains auteurs soulignent

[74] Bronner, *Apocalypse cognitive*.

[75] Maik Larooij et Petter Törnberg, « Can We Fix Social Media? Testing Prosocial Interventions Using Generative Social Simulation », arXiv:2508.03385, prépublication, arXiv, 5 août 2025, doi:10.48550/arXiv.2508.03385.

[76] Andreas Matthias, « The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata », *Ethics and Information Technology*, 6.3 (2004), pp. 175–83, doi:10.1007/s10676-004-3422-1.

qu'une question importante en arrière-plan est celle du contrôle des systèmes d'IA. « Le véritable problème se situe ailleurs : les machines autonomes devraient être conçues de manière à présenter un niveau de risque moralement acceptable. Si elles ne répondent pas à cette norme, elles présentent ce que nous appelons un « déficit de contrôle ». »[77] Au-delà de la possibilité d'attribuer la responsabilité, il est primordial de garantir un « contrôle humain significatif ».[78]

Il est donc primordial de promouvoir la culture de l'IA afin de développer les capacités des personnes et des communautés à évaluer correctement le type d'outils qu'elles mobilisent. Ce qui importe, c'est de savoir à quel système on peut faire confiance pour quoi, afin de distinguer correctement les cas d'utilisation valables et fructueux de ceux qui sont dangereux. L'automatisation ou la délégation importante ne devrait avoir lieu qu'avec des systèmes auxquels on peut faire confiance, soit parce que les utilisateurs finaux les ont testés eux-mêmes, soit (plus probablement) lorsqu'ils confient l'évaluation à des sous-groupes. Les systèmes auxquels on ne peut pas faire suffisamment confiance (tels que l'IA générative et les LLM) doivent toujours être utilisés sous supervision humaine directe ou dans des contextes où des résultats invalides ne posent pas de problème. Le même type de capacités est nécessaire pour contribuer à décider des outils et des technologies que nous devrions développer à l'avenir. Il est par exemple crucial que des communautés élargies participent à la réflexion sur le problème spécifique de la perte de contrôle sur les systèmes d'IA les plus avancés. Pour y parvenir, il faut aller au-delà de la prophétie habituelle selon laquelle l'IA deviendrait plus intelligente que les humains. Comme nous l'avons dit, le problème consiste plutôt à s'assurer que des systèmes mécaniques complexes et puissants ne deviennent pas trop imprévisibles et en décalage avec nos objectifs et nos valeurs (cela concerne directement la question du « contrôle humain significatif » abordée ici). Il s'agit également de ne pas déployer de systèmes imprévisibles dans des contextes critiques. Dans cette perspective, le débat devrait davantage porter sur la nouvelle tendance de l'« IA agentive », où l'IA générative (LLM ou similaire) ne se limite plus à la production de texte (ou de contenu similaire) mais peut exécuter un éventail élargi d'actions (piloter d'autres programmes pour automatiser la gestion de l'agenda et la prise de rendez-vous, les achats, la messagerie…), jusqu'à la compilation et l'exécution de code. Dans de telles nouvelles configurations, les conséquences d'un dysfonctionnement et d'un désalignement peuvent devenir extrêmement dangereuses.

Plus globalement, garantir un contrôle humain significatif dépendra toujours de la capacité des différents acteurs concernés à exercer leur esprit critique et leur capacité de prise de décision au sens fort (fondée sur la possibilité de contourner, d'être confronté à et d'arbitrer entre une pluralité d'options…). On voit ici réapparaître un schéma important dans la perspective de l'éthique et de la réglementation de l'IA : la technologie de l'IA peut fortement influencer

[77] Frank Hindriks et Herman Veluwenkamp, « The Risks of Autonomous Machines: From Responsibility Gaps to Control Gaps », *Synthese*, 201.1 (2023), p. 21, doi:10.1007/s11229-022-04001-5.

[78] Filippo Santoni de Sio et Giulio Mecacci, « Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them », *Philosophy & Technology*, 34.4 (2021), pp. 1057–84, doi:10.1007/s13347-021-00450-x.

(positivement ou non) les capacités humaines qui sont essentielles pour garantir son propre développement et son utilisation adéquats.

Un élément est particulièrement préoccupant à cet égard. Il semble de plus en plus évident que la technologie de l'IA s'accompagne d'une tentation de se décharger du fardeau d'agir et de décider.[79] D'une part, cette tentation peut être liée à une sorte de honte prométhéenne[80] qui conduit les gens à penser que les systèmes d'IA sont plus capables qu'eux-mêmes. Une étude empirique récente a mis en évidence que les gens ont tendance à s'en remettre davantage à l'IA (générative) lorsqu'ils ne se sentent pas sûrs d'eux.[81] Dans ce contexte, il faut s'interroger sur la place que nous voulons accorder à la prise de risque et, par conséquent, aux erreurs humaines. Évidemment, nous devons nous efforcer de limiter nos erreurs et leurs conséquences. Mais nous ne devons pas laisser notre discernement être entièrement guidé par une sorte d'aversion aveugle et monolithique pour les erreurs. Le risque existe même de déléguer certaines tâches à l'IA alors même que nous percevons certaines limites de l'outil, simplement parce que nous voulons éviter d'assumer la responsabilité de nos actes. Il peut être tentant de ne pas s'opposer à un résultat issu de l'IA même lorsqu'il semble problématique, car, en cas de problème, « c'est la faute de la machine ». Ce risque est particulièrement aigu pour les prises de décision à enjeux élevés, comme dans le contexte médical. Il est également intéressant de rappeler qu'un tel transfert de responsabilité est incompatible avec la préservation de la créativité humaine et de la marge de manœuvre nécessaire pour accompagner l'utilisation de systèmes algorithmiques automatisant la prise de décision par le biais d'analyses prédictives basées sur des données passées. Seul un tel accompagnement autonome et créatif peut garantir que les personnes ne soient pas soumises à un traitement algorithmique qui nie leur propre nature créative, leur propre possibilité de contourner et de modifier leur manière d'agir et de vivre.

Dans l'ensemble, le délestage peut répondre à la tentation de se réduire à un objet inerte,[82] un être purement cybernétique s'efforçant de minimiser les efforts qu'il déploie en quête de satisfaction libidinale.[83] En fait, la capacité à contourner les obstacles, à imaginer que les choses pourraient être différentes de ce qu'elles sont réellement et à tenter d'influencer le cours des événements est au cœur de ce que signifie être humain. Mais c'est une capacité extrêmement exigeante, que nous devrions cultiver et protéger, peut-être par le biais d'une « déclaration universelle des droits de l'esprit humain ».[84] Plus qu'un risque de perte soudaine de contrôle

[79] Evan F. Risko et Sam J. Gilbert, « Cognitive Offloading », *Trends in Cognitive Sciences*, 20.9 (2016), pp. 676–88, doi:10.1016/j.tics.2016.07.002 ; Michael Gerlich, « AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking », *Societies*, 15.1 (2025), p. 6, doi:10.3390/soc15010006.

[80] Günthe Anders, Günther/Dries Anders Christia et Christopher John Müller, *The Obsolescence of the Human* (University of Minnesota Press ; Univ Of Minnesota Press, s.d.).

[81] Hao-Ping (Hank) Lee et al., « The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers », *Actes de la conférence CHI 2025 sur les facteurs humains dans les systèmes informatiques* (New York, NY, États-Unis), CHI '25, 25 avril 2025, pp. 1–22, doi:10.1145/3706598.3713778.

[82] Jean-Michel Besnier, *N'être plus qu'un objet : la tentation d'oublier la vie*, Technologia (Hermann éditeurs des sciences et des arts, 2025).

[83] Mark Hunyadi, « La bataille de l'esprit », *Esprit*, n° 4 (avril 2025), pp. 43–53, doi:10.3917/espri.2504.0043.

[84] Mark Hunyadi, *Déclaration universelle des droits de l'esprit humain : une proposition* (PUF, 2024).

sur l'IA, comme dans « les scénarios de prise de pouvoir brutale couramment évoqués dans le domaine de la sécurité de l'IA », l'humanité pourrait bien être confrontée au danger d'une « perte progressive de pouvoir », d'une « érosion graduelle de l'influence humaine » qui pourrait conduire à une « perte irréversible de l'influence humaine sur des systèmes sociétaux cruciaux, précipitant une catastrophe existentielle par la perte permanente de pouvoir de l'humanité ».[85]

Une fois encore, il faudra beaucoup d'efforts, d'intelligence et d'autonomie de la part des humains pour orienter le développement et l'utilisation de systèmes d'IA qui nous soutiendront et nous donneront les moyens d'agir dans nos activités décisionnelles, qui préserveront et favoriseront notre autonomie et nos capacités décisionnelles. Comment encourager le développement d'algorithmes qui limitent la tentation d'un transfert illégitime de responsabilités ? Quel type de prise de décision souhaitons-nous déléguer aux machines, sachant que cela reviendra en fin de compte à remplacer la prise de décision par l'automatisation ? Quelle est la place des assistants et conseillers IA dans nos processus de prise de décision (morale) ? Quelle place accordons-nous à l'erreur humaine ? Quel prix sommes-nous prêts à payer pour défendre, voire cultiver, l'autonomie humaine ? Ce sont là des questions très pointues qui devraient être débattues au sein de communautés aussi larges que possible de personnes concernées.

| Problématiser les notions de progrès, de bonne vie et de vulnérabilité (recommandation 11) |
|---|

En toile de fond de la discussion développée jusqu'à présent se trouve la question clé de la bonne vie. Que signifie mener une bonne vie ? Que signifie améliorer nos vies ? Quel est le lien avec les idées de progrès et d'innovation, en particulier avec la technologie de l'IA ? Sur ces questions, on peut difficilement adopter sans réserve la position des techno-optimistes ou des techno-solutionnistes radicaux, telle que défendue par certains acteurs puissants de la Silicon Valley comme Peter Thiel ou Marc Andreessen, selon laquelle : « il n'existe aucun problème matériel – qu'il soit créé par la nature ou par la technologie – qui ne puisse être résolu par davantage de technologie » ou que « nous sommes sur le point d'un décollage de l'intelligence qui étendra nos capacités à des sommets inimaginables », l'IA étant considérée comme « notre alchimie, notre pierre philosophale – nous faisons littéralement penser le sable ».[86] Ce genre de discours présuppose un lien de causalité fondamental entre l'innovation et le progrès technologique d'une part, et une véritable amélioration de nos vies d'autre part.

Pour orienter de manière adéquate le développement et l'utilisation de l'IA, il est nécessaire de résister à ce type de raccourcis techno-solutionnistes. Il est particulièrement important d'être plutôt toujours prêt à affiner et à élargir notre compréhension de ce que signifie le « véritable progrès humain ». Avant de rechercher des solutions technologiques à un problème donné, nous devons nous assurer que notre analyse de celui-ci est suffisamment approfondie et riche.

---

[85] Jan Kulveit et al., « Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development », arXiv:2501.16946, prépublication, arXiv, 29 janvier 2025, doi:10.48550/arXiv.2501.16946.

[86] Andreessen, « The Techno-Optimist Manifesto » ; voir également l'interview de Peter Thiel pour le New York Times : Ross Douthat, « Opinion | Peter Thiel and the Antichrist », Opinion, *The New York Times*, 26 juin 2025 <https://www.nytimes.com/2025/06/26/opinion/peter-thiel-antichrist-ross-douthat.html> [consulté le 24 octobre 2025].

Par exemple, il est loin d'être évident que l'on puisse répondre de manière adéquate à l'épuisement extrême des professionnels de santé en se contentant de leur fournir des outils améliorant leur efficacité et leur productivité, ou en les remplaçant pour certaines tâches par des robots ou d'autres automates. Ces solutions ne peuvent être légitimes qu'en tant que résultats d'une large réflexion collective (tenant compte des réalités des pratiques sur le terrain) sur les causes profondes de la surcharge de travail et du manque de personnel. De même, il ne suffit pas de répondre au sentiment croissant de solitude qui touche un nombre toujours plus grand de personnes par des compagnons artificiels et la robotique sociale. Dans de tels cas, la solution technologique peut même devenir un moyen de perpétuer, voire d'aggraver, un problème plus profond.

Une telle prudence et une telle finesse dans l'analyse du lien entre, d'une part, le progrès technologique et l'innovation et, d'autre part, le véritable progrès humain sont au cœur de la tradition technocritique que nous avons évoquée dans l'introduction. On pourrait facilement associer à cette école de pensée les paroles du pape François mettant en garde contre notre tendance à faire un usage (moralement) aveugle de nos pouvoirs (technologiques), en partant du principe que plus de pouvoir équivaut nécessairement au progrès humain (réduisant souvent notre compréhension de la notion à des questions d'utilité ou de sécurité présentées comme des impératifs).[87] Cette tendance culmine dans ce que François appelle le « paradigme technocratique » qui « exalte le concept d'un sujet qui, à l'aide de procédures logiques et rationnelles, s'approche progressivement d'un objet extérieur et en prend le contrôle » dans un esprit de « possession, maîtrise et transformation ». Avec les paradigmes technocratiques, les humains ont tendance à considérer la réalité dans laquelle ils interviennent comme « quelque chose d'informe, totalement ouvert à la manipulation », dont il faut extraire le plus possible (au lieu d'« être en harmonie avec les possibilités offertes par les choses elles-mêmes et de les respecter »).[88] Dans un tel environnement de pensée marqué par la domination de la rationalité instrumentale, il devient difficile de concevoir le progrès autrement qu'en termes d'efficacité et de performance mesurable. Les problèmes que les humains s'efforcent de résoudre, et même les affaires humaines en général, ont tendance à être réduits à des indicateurs à optimiser. Les humains eux-mêmes deviennent de simples ressources, compétences et rôles ou fonctions, « qui peuvent alors être reproduits, améliorés, surpassés ».[89]

Non seulement ce type de mentalité peut conduire à promouvoir ou à privilégier des formes mécanistes et algorithmiques de gouvernance,[90] mais il nourrit également le risque de mutiler la quête humaine légitime de liberté et d'émancipation, en la réduisant à un rejet obsessionnel de toute limite. Dans cette perspective, toute faiblesse, toute vulnérabilité, toute possibilité

---

[87] Pape François, « Laudato Si' (24 mai 2015) », 24 mai 2015, par. 105 <https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html> [consulté le 5 décembre 2025].

[88] Pape François, « Laudato Si' (24 mai 2015) », par. 106.

[89] Notre traduction, Nicolas Léger et Adrien Tallent, « L'IA aux frontières de l'esprit. Introduction », *Esprit*, n° 4 (avril 2025), p. 35-42 (p. 36), doi:10.3917/espri.2504.0035.

[90] Antoinette Rouvroy, « Algorithmic Governmentality », dans *More Posthuman Glossary*, sous la direction de Rosi Braidotti, Emily Jones et Goda Klumbytė (Bloomsbury Academic, 2022), doi:10.5040/9781350231467.

d'échec est un défaut qu'il faut s'efforcer de corriger. Une telle vision du développement humain comme progression vers la perfection et une puissance illimitée est hautement problématique pour de multiples raisons, dont la moindre n'est pas sa dimension élitiste et inégalitaire.[91] Nous souhaitons ici nous concentrer sur les raisons épistémologiques, morales et anthropologiques de prendre ses distances par rapport à cette conception mutilée du progrès humain.

Comme nous l'avons vu dans les sections précédentes (recommandations 8 à 10), la possibilité même de l'erreur et de l'échec est intrinsèque à la connaissance et à la prise de décision. Tenter de réduire le nombre d'erreurs que nous commettons est un devoir et fait partie intégrante de la connaissance et de la prise de décision morale. Mais affirmer que nous devrions lutter contre la possibilité même de l'échec et de l'erreur est tout autre chose : cela ne revient pas à améliorer l'élaboration de la connaissance et la prise de décision, mais à les éradiquer. Comme nous l'avons vu, la connaissance et la prise de décision au sens fort impliquent une véritable marge de manœuvre incluant de manière irréductible la possibilité de faire un mauvais choix. Améliorer nos connaissances et nos décisions ne nécessite pas de réduire cette marge de manœuvre dans une quête illusoire d'infaillibilité. Cela exige plutôt d'affiner notre sens collectif des responsabilités, notre capacité à exercer notre esprit critique et notre sens commun. S'améliorer dans ce domaine est un chemin de vie, un engagement à faire de notre mieux (individuellement et collectivement) pour décider du mieux que nous pouvons sans garantie absolue, pour agir et croire pour de bonnes raisons (bien que faillibles).

On peut en dire autant de la vulnérabilité dans un sens plus large et encore plus fondamental. La robustesse absolue du corps et de l'esprit ne constitue pas la perfection de la vie humaine, mais sa négation radicale. Là encore, il est de notre devoir et de notre responsabilité de faire de notre mieux pour soigner ou prévenir les blessures et les maladies. Il est également tout à fait légitime d'essayer d'éviter de blesser les autres et d'être blessé par eux. Mais cela n'équivaut pas à essayer d'éradiquer autant que possible la possibilité d'être blessé, de tomber malade ou même de mourir. Nous pouvons souffrir et être blessés parce que nous sommes vulnérables. Mais ce serait une erreur dangereuse de réduire la vulnérabilité à ces seuls aspects négatifs. Comme le dit David Doat, « [l]a vulnérabilité n'est ni une faiblesse ni une pauvreté. Elle ne peut pas non plus être réduite à la vieillesse, au handicap ou à la maladie. (…) nous devons distinguer entre « vulnérabilité » et « vulnération ». La première fait référence à la possibilité d'être affecté dans sa structure physique ou psychologique ; la seconde fait référence à l'état qui suit une blessure. Il est important de faire la distinction. Lors d'une rencontre amoureuse, par exemple, les amants sont dans un état de vulnérabilité car ils s'exposent l'un à l'autre, chacun se laissant affecter par l'être aimé, mais aucun des deux n'est blessé. »[92]

En résumé, la notion de progrès humain authentique ne peut se réduire à un accroissement de pouvoir qui nous rapprocherait de la perfection, de l'infaillibilité et de l'invulnérabilité. Tout ce

[91] Michael J. Sandel, *The Case against Perfection: Ethics in the Age of Genetic Engineering* (Harvard University Press, 2009), doi:10.2307/j.ctvjz80mc.

[92] Extrait de Brigitte Bègue, « La vulnérabilité peut être une chance. Mais on l'oublie », entretien avec David Doat, 5 mars 2021, Actualités sociales hebdomadaires N° 3199 <https://www.ash.tm.fr/hebdo/3199/entretien/la-vulnerabilite-peut-etre-une-chance-mais-on-loublie-634607.php> [consulté le 5 décembre 2025], notre traduction.

qui compte dans nos vies se paie au prix de la vulnérabilité et de la faillibilité. C'est parce que nous avons la capacité de choisir que nous pouvons commettre des erreurs. C'est parce que nous pouvons juger de la qualité des raisons de croire en quelque chose et imaginer des alternatives que nous pouvons élaborer des connaissances, mais cela implique de manière irréductible la possibilité d'évaluations erronées. C'est parce que nous sommes vivants que nous pouvons être affectés négativement, blessés et traumatisés. C'est parce que nous pouvons ressentir de la joie et de l'amour que nous pouvons aussi ressentir de la tristesse et du désespoir. Se demander si et comment l'IA peut nous rendre tous puissants, infaillibles et invulnérables est donc une mauvaise question. Nous devrions plutôt réfléchir à la manière de développer et d'utiliser des systèmes d'IA qui pourraient nous aider à mieux apprivoiser et équilibrer la vulnérabilité ambivalente mais essentielle qui réside au plus profond de ce que nous sommes.

### Cultiver notre sensibilité à la vie et à l'expérience vécue en pleine conscience (recommandation 12)

Ce que nous venons de dire nous amène à un dernier sujet que nous souhaitons explorer dans la perspective de renforcer notre capacité à orienter correctement le développement et l'utilisation de l'IA. Il s'agit de la légitimité et de l'importance de valoriser la vie et l'expérience vécue, au cœur desquelles se trouvent l'affectabilité et la vulnérabilité. Ces dimensions sont essentielles à des relations authentiques entre les humains et, plus largement, entre les êtres vivants. À cet égard, nous devons à nouveau mettre en garde contre l' e la tendance « outcomiste » qui tend à écarter ces dimensions au motif qu'elles ne sont ni objectivables ni démontrables, nous incitant à nous concentrer uniquement sur les résultats. Bien que nous ayons consacré plusieurs sections à discuter et à désamorcer ce raisonnement (voir recommandation 7), ses conséquences potentiellement néfastes sur notre capacité à construire une technologie d'IA véritablement au service de l'humanité.

S'appuyant sur l'outcomisme, certains défendent la possibilité d'une conscience de l'IA (dans un avenir prévisible) ou, de manière plus pernicieuse, affirment qu'il est vain d'essayer de différencier les humains et les machines sur cet aspect. Il est inutile de se demander si l'IA est véritablement consciente au-delà de ses comportements et résultats observables. Si nous avons le sentiment qu'elle est consciente parce qu'elle se comporte de manière convaincante, alors elle est consciente… il n'y a rien de plus à dire. L'important, c'est ce qu'elle nous fait ressentir.

Selon ce raisonnement, de nombreuses utilisations de l'IA dans les relations sociales deviennent plus facilement acceptables. Certaines données empiriques suggèrent que les personnes, lorsqu'on leur présente en aveugle des communications médicales écrites produites par des modèles de langage (LLM) ou par de véritables professionnels de santé, ont tendance à préférer celles produites par la machine, notamment en raison de leur contenu plus empathique.[93] Si les gens préfèrent ces réponses, pourquoi ne pas leur donner ce qu'ils

[93] L'expérience menée à partir des réponses à des questions posées sur des forums consacrés à la santé. Voir : John W. Ayers et al., « Comparing Physician and Artificial Intelligence Chatbot Responses to Patient

préfèrent ? De même, si les gens aiment les œuvres d'art générées par l'IA (du moins lorsque nous n'insistons pas trop sur leur origine), pourquoi devrions-nous nous abstenir de produire de la musique, des images ou d'autres contenus de cette manière ? Si les gens considèrent les compagnons IA comme de véritables amis capables de leur apporter des relations sociales et un réel réconfort, pourquoi les priver de ce bien-être effectif ?

Comme nous l'avons vu, il n'y a aucune raison impérieuse d'adopter l'outcomisme. Au contraire, il existe de nombreuses façons d'argumenter en faveur de nos intuitions fondamentales et traditionnelles sur le type d'entités qui sont vivantes ou non, qui sont capables de conscience et d'affectabilité au sens psychologique, ou qui peuvent entrer dans des relations authentiques. Cela signifie qu'il est tout à fait légitime de valoriser non seulement la qualité des résultats qui nous sont présentés, mais aussi la manière dont ils ont été produits, et en particulier la présence d'une expérience vécue, de la vulnérabilité et de l'affectabilité, la présence d'une personne authentique en amont. Pour orienter correctement le développement et les utilisations des systèmes d'IA, nous devons donc cultiver notre sensibilité à la vie et aux personnes véritablement vulnérables et affectables. Plus précisément, nous devons favoriser et cultiver notre capacité à valoriser leur présence (et pas seulement l'apparence d'une telle présence), à évaluer quand leur présence a une valeur décisive.

Un premier domaine dans lequel la présence d'une expérience vécue authentique est indispensable est celui du rapport à la connaissance et à la vérité, ainsi que celui de la prise de décision. Il n'y a pas d'élaboration de la connaissance ni de prise de décision au sens fort sans qu'une ou plusieurs personnes authentiques fassent un usage plus ou moins responsable de la liberté dont elles disposent au sein de leur expérience vécue. Dans de nombreux autres domaines, savoir qu'il y a quelqu'un en face de nous (plutôt que de n'avoir que de simples apparences) est essentiel. C'est particulièrement le cas dans le domaine de la santé, où l'empathie et les relations authentiques entre médecin et patient jouent un rôle crucial pour la thérapie et le rétablissement.[94] On pourrait en dire autant de l'importance des relations humaines dans les processus psychothérapeutiques. La présence d'une personne authentique à qui nous parlons lors des conversations thérapeutiques est indispensable,[95] même si cela peut sembler plus exigeant que de simplement parler à un chatbot.

En effet, les relations humaines authentiques nous exposent à des risques et s'accompagnent d'une part d'inconfort, d'insécurité et d'inquiétude. Il existe un risque spécifique lié à l'IA sociale et aux compagnons artificiels. Dans la plupart des cas, les systèmes d'IA de ce type sont conçus comme des produits à notre service. Ils ne s'opposeront jamais fermement à notre volonté. Contrairement à l'angoisse que peuvent susciter les relations authentiques, il peut sembler très attrayant d'obtenir une apparence très convaincante de relation, mais débarrassée de tout risque d'être rejeté, abandonné, jugé ou blessé. Aussi tentant que cela puisse être, cela reviendrait néanmoins à supprimer tout ce qui fait une véritable relation humaine, à vider de

Questions Posted to a Public Social Media Forum », *JAMA Internal Medicine*, 183.6 (2023), pp. 589–96, doi:10.1001/jamainternmed.2023.1838.

[94] Joan C. Tronto, *Caring Democracy: Markets, Equality, and Justice* (New York University Press, 2013).

[95] Jana Sedlakova et Manuel Trachsel, « L'intelligence artificielle conversationnelle en psychothérapie : un nouvel outil thérapeutique ou un agent ? », *The American Journal of Bioethics*, 23.5 (2023), pp. 4–13, doi:10.1080/15265161.2022.2048739.

toute valeur cette imitation d'acceptation et d'amour. Comment cela pourrait-il être véritablement de l'amour ou de l'acceptation s'il ne s'inscrit pas dans le contexte d'une expérience vécue authentique par une personne libre qui ne pouvait pas (n'aurait pas pu) l'offrir ?[96] Des données empiriques semblent confirmer ces limites des compagnons IA lorsqu'ils sont utilisés à grande échelle pour compenser l'isolement social.[97]

La valeur de la présence d'une personne réelle devient évidente lorsqu'on considère l'expression de mots de compassion, par exemple face à une personne mourante. Les mots eux-mêmes (ou d'autres moyens d'expression) ne sont pas le plus important dans une telle situation. Ce qui prime, c'est qu'ils traduisent un sentiment de compassion sincèrement ressenti. On pourrait dire la même chose dans le cas d'un enfant qui nous fait un dessin. Dans ce contexte, la qualité esthétique « objective » du résultat (le dessin) est en soi largement secondaire. Quelle serait la valeur d'un dessin objectivement très beau que l'enfant nous offre, mais qui aurait été produit en quelques minutes par un programme d'IA générative ? Pourrait-on la comparer à la valeur d'un dessin imparfait (dont on ne sait même pas vraiment ce qu'il représente) que l'enfant a mis plusieurs heures à réaliser, y mettant beaucoup d'efforts et d'intention ? Nous ne nous soucions pas vraiment ici de la qualité esthétique « objective » de l'image. Ce qui importe, c'est l'expérience vécue intense dont elle est le fruit.

D'une certaine manière, les choses peuvent se situer quelque part entre les deux dans le cas de l'art « professionnel ». L'aspect « objectif » de l'œuvre d'art peut légitimement avoir son importance. Dans la plupart des cas, nous ne prenons pas en compte uniquement ou principalement l'expérience vécue par l'artiste lorsqu'il ou elle a élaboré l'œuvre. Mais nous prenons également en compte ces dimensions. Nous sommes en droit d'intégrer dans notre jugement esthétique notre connaissance de l'origine de l'œuvre d'art, de son auteur. Plus encore, il est parfaitement légitime que le fait de savoir qu'elle a été réalisée par une personne réelle plutôt que par un système d'IA nous fasse percevoir l'œuvre différemment. Remettre en cause ou nier cette légitimité (en particulier sur la base d'arguments épistémologiques erronés concernant l'obligation d'être objectif) constitue une grave atteinte aux dimensions fondamentales de ce que signifie être vivant et être un être huma . Nous devons apprendre à opposer une forte résistance à une telle agression en (répétons-le) favorisant et cultivant notre sensibilité à la vie et à la présence humaine authentique.

Forts d'un tel bagage, nous pouvons nous interroger sur les types de systèmes d'IA et d'utilisations souhaitables, à savoir ceux qui contribueront à développer notre sensibilité à la vie et à cultiver notre capacité à nouer des relations authentiques. D'une manière générale, il est indispensable de toujours indiquer lorsqu'un résultat est produit par un système d'IA plutôt que par un être humain. Comme nous venons de le voir, cette information est nécessaire pour que les gens puissent évaluer ce qui leur est présenté. Au-delà de cela, nous devons déployer des efforts sérieux pour identifier les utilisations véritablement positives de systèmes capables

[96] Groupe de recherche sur l'IA du Centre for Digital Culture, « Encountering Artificial Intelligence », p. 117.

[97] Cathy Mengying Fang et al., *Comment l'IA et les comportements humains façonnent les effets psychosociaux de l'utilisation des chatbots : une étude longitudinale contrôlée* (MIT Media Lab et Open AI, 2025) <https://www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudinal-controlled-study/> [consulté le 29 novembre 2025].

d'imiter de manière convaincante les humains ou d'autres êtres vivants, ainsi que pour déterminer les caractéristiques que ces systèmes devraient ou ne devraient pas posséder. C'est une question très difficile.

Prenons par exemple le cas des compagnons IA. On peut considérer qu'on pourrait les utiliser comme des jouets sophistiqués ou comme un support de divertissement, tant qu'on sait qu'ils ne sont pas humains. Après tout, les compagnons IA peuvent être comparés à un bon film, un bon livre, ou, mieux encore, à un bon jeu vidéo dont on aime suivre les aventures des personnages et avec lesquels on interagit éventuellement. Il pourrait y avoir là des utilisations véritablement positives ou inoffensives. Il faut néanmoins prêter attention à un point délicat : il ne suffit pas toujours d'être simplement clairement conscient que l'on interagit avec un système d'IA imitant l'humain et non avec une personne réelle. En effet, les compagnons IA poussent à l'extrême l'attribution d'une apparence humaine à des artefacts conçus pour servir les utilisateurs et les consommateurs. Les systèmes d'IA atteignant des niveaux convaincants en la matière augmentent le risque « d'éduquer leurs utilisateurs à la négation de l'autre et de favoriser une culture qui absorbe l'intimité dans un schéma de relations de propriété et de droits plutôt que dans le don vulnérable de la véritable intersubjectivité ». « Là où il n'y a pas d'« autre », mais seulement *l'apparence* d'un autre à notre disposition, parallèlement à l'absence de l'exigence qui s'exercerait sur le don de soi par la confrontation avec un véritable autre, nous risquons d'être conditionnés à un talent dangereux pour l'exploitation. »[98]

Nous risquons ainsi de favoriser la réduction progressive des relations humaines à des interactions de service, avec le danger de devenir de moins en moins capables de tolérer la véritable autonomie des autres, cette autonomie qui rend les relations authentiques. Au lieu d'être capables de voir les frictions et l'opposition comme autant d'occasions de rencontre authentique avec de vraies personnes, nous pourrions commencer à considérer les personnes responsables de ces résistances comme des êtres humains défectueux (comme nous le ferions avec des artefacts défectueux). Ces questions pourraient ouvrir un débat parallèle aux discussions sur le bien-être et les droits de l'IA, mais pour des raisons différentes. En effet, il pourrait devenir nécessaire de réglementer la manière dont nous interagissons avec des compagnons ou assistants artificiels, non pas parce que nous pourrions leur nuire par des comportements inappropriés, mais parce que nous pourrions nous nuire à nous-mêmes.[99]

Nous pouvons ainsi constater qu'il faudra beaucoup de réflexion, d'exploration et d'efforts pour faire place à la vie et à l'expérience vécue dans les contextes où leur valeur est primordiale. S'il peut être facile et tentant d'utiliser les systèmes d'IA pour proposer des substituts (pâles) aux interactions sociales, mettre l'IA au service de l'intensification des relations authentiques et de leur qualité constitue un défi exigeant et d' . Les technologies numériques (en particulier les réseaux sociaux et désormais l'IA sociale) peuvent donner une impression superficielle de

[98] Groupe de recherche sur l'IA du Centre pour la culture numérique, « Rencontre avec l'intelligence artificielle », p. 120.

[99] Groupe de recherche sur l'IA du Centre for Digital Culture, « Encountering Artificial Intelligence », p. 128-130.

proximité sociale alors qu'elles contribuent en réalité à l'isolement et à l'aliénation.[100] Des risques similaires peuvent se présenter selon la manière dont nous réagirons aux résultats empiriques démontrant l'impression d'empathie que l'IA générative peut susciter. Elles semblent parfois plus empathiques que les humains, comme nous l'avons mentionné à propos des réponses sur les forums publics consacrés à la santé. Bien sûr, les auteurs ne concluent pas que les professionnels de santé devraient être remplacés par des systèmes d'IA. Ils suggèrent plutôt que les LLM pourraient rédiger des supports de communication plus empathiques pour ces professionnels (avec la possibilité d'améliorer les parcours thérapeutiques des patients et de réduire la surcharge de travail des professionnels).[101] Plus globalement, il est souvent proposé, en particulier dans ce domaine des soins de santé, d'utiliser des systèmes d'IA et de la robotique avancée pour soulager les professionnels surchargés.

À notre avis, cette exploration peut être utile, mais elle ne devrait pas être menée sans une analyse approfondie des causes de la surcharge de travail ou d'autres problèmes tels que les difficultés de communication des professionnels (de la santé) (en termes purement médicaux et en matière d'empathie). Nous devons rechercher les causes de ces problèmes et identifier les différentes options pour les atténuer. Il est loin d'être évident que le recours à l'IA soit la meilleure option. En particulier, externaliser la production d'une communication apparemment empathique pourrait même aggraver l'épuisement des professionnels de santé. Peut-être aimeraient-ils consacrer plus de temps à établir de véritables relations avec leurs patients, mais ne le peuvent pas en raison de la surcharge de travail. Peut-être faudrait-il également s'interroger sur leur formation, notamment en ce qui concerne une éventuelle tendance à réduire les êtres vivants et les personnes humaines à des fonctions biologiques à surveiller, maintenir ou restaurer. Là encore, il est primordial de réfléchir en profondeur aux véritables objectifs que les systèmes d'IA devraient servir, en s'efforçant de placer au centre de nos efforts de discernement des dimensions clés telles que la valeur de la vie et l'expérience de vie consciente dont nous avons discuté dans cette section.

## Remarques finales

Terminons cette exposition par quelques remarques générales et en soulignant les messages les plus importants. Avant tout, il est important d'insister sur l'esprit de l'effort collectif de discernement que nous proposons ici. En aucun cas cela ne peut se résumer à affirmer simplement que l'humain est toujours merveilleux et doit être préféré à la machine dans tous les contextes. Bien sûr, les humains sont faillibles, ils commettent des erreurs, ils peuvent se montrer particulièrement désagréables les uns envers les autres. Certains humains peuvent se comporter de manière totalement barbare et inhumaine. Notre argument central est plutôt de dire que la meilleure façon d'avancer n'est pas toujours de s'efforcer de faire en sorte que les systèmes d'IA assurent des fonctions où les humains peuvent se tromper, surtout lorsque cela implique une illusion d'infaillibilité ou une simple apparence de gentillesse et de bienveillance.

[100] Sherry Turkle, *Seuls ensemble : pourquoi nous attendons plus de la technologie et moins les uns des autres*, troisième édition, édition de poche révisée (Basic Books, 2017).

[101] Ayers et al., « Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum ».

La meilleure façon d'avancer ne peut être de mettre les humains et les machines en concurrence. Comme nous l'avons précisé, il est la plupart du temps illégitime de prétendre que l'IA peut remplacer les humains dans certaines tâches. Ce qui est vrai, c'est que l'IA peut assurer certaines fonctions qui exigeraient autrement une présence humaine. Mais cela signifie rarement que le remplacement de l'humain par l'IA, en tant qu' , aboutit à une situation équivalente. Nous devons toujours garder à l'esprit que nous remplaçons une personne vivante, avec toute la richesse de son expérience vécue, par un pur mécanisme et un automatisme.

En résumé, la question principale n'est pas de déterminer ce que l'IA peut faire mieux que les humains (même si c'est une question légitime lorsqu'on l'aborde dans toute sa complexité, sans réduire la manière dont nous comprenons les situations). La question principale dont nous devrions partir est plutôt : « comment la technologie de l'IA peut-elle nous aider à devenir de meilleurs humains ? ». C'est un changement de perspective essentiel à adopter si nous voulons mettre en place une éthique et une réglementation solides en matière d'IA. La question concernant l'IA ne devrait pas être : « quelle est notre place en tant qu'êtres humains dans le nouveau monde de l'IA ? », comme si la technologie de l'IA n'était pas le résultat des choix et des actions de (certains) humains et que la seule chose que les communautés humaines pouvaient faire était de s'adapter à ce nouveau monde. Nous devrions plutôt nous demander collectivement : « quelle est la place de la technologie de l'IA dans notre monde humain, et par rapport aux autres êtres vivants ? » « Quelle contribution l'IA peut-elle apporter au développement de sociétés plus humaines ? ». Formulée ainsi, il apparaît clairement que la question du développement et de l'utilisation de la technologie de l'IA est (au moins) autant un enjeu politique et éthique qu'une question technique.

En réalité, nous ne pouvons répondre à ces questions sans nous interroger collectivement sur ce que signifie devenir de meilleurs êtres humains, et sur les projets de société auxquels l'IA devrait se mettre au service. Cela exige un engagement politique considérable de la part d'une grande partie des communautés humaines. Nous devons réfléchir ensemble, faire appel à notre esprit critique et à notre capacité à prendre de véritables décisions. À cet égard, il est important de rappeler la menace circulaire que fait peser l'IA. Nous avons besoin de beaucoup de temps d'attention libre et de bonne qualité pour participer de manière adéquate à ces efforts de discernement, alors que les algorithmes de recommandation sont actuellement très efficaces pour détourner ce temps d'attention libre. Nous devons cultiver notre capacité à prendre des décisions au sens fort, alors que l'IA générative ouvre des possibilités tentantes d'externalisation cognitive excessive. Nous devons cultiver notre sensibilité à la vie et à l'importance de l'expérience phénoménale véritablement vécue à une époque où les machines imitent de mieux en mieux. Sans un discernement suffisant, les systèmes d'IA développés et les utilisations adoptées sans esprit critique peuvent saper les capacités dont nous avons besoin pour discerner correctement et construire une technologie d'IA souhaitable. Pour nourrir ces efforts de discernement, nous devons favoriser une meilleure compréhension tant de la technologie d'IA elle-même que de la nature et de la condition humaines. La culture de l'IA est absolument essentielle pour saisir correctement ce que nous pouvons raisonnablement attendre ou non des systèmes d'IA (en particulier en termes de fiabilité). De plus, nous devons toujours approfondir le contexte d'une difficulté que nous rencontrons et que nous

envisageons d'atténuer grâce à l'IA. Nous devons toujours nous demander si l'IA apportera une véritable solution au problème ou si elle peut, au mieux, nous aider à faire face temporairement à un problème plus profond que nous devrions aborder différemment. Ne pas le faire sérieusement reviendrait à prendre le risque d'utiliser l'IA pour perpétuer des difficultés humaines aiguës, nous privant ainsi de la possibilité de mieux nous développer et de nous épanouir.

Encore une fois, il est légitime d'être enthousiaste et optimiste à l'égard de l'IA. Nous devons l'être. L'IA recèle un potentiel énorme pour nous soutenir dans notre développement et notre épanouissement. Mais nous devons en même temps toujours cultiver et garder clairement à l'esprit le prix que nous sommes appelés à payer en termes d'engagement dans des efforts exigeants de discernement éthique et politique.