

JÉRÔME BÉRANGER

PRÉFACE DE GILLES BABINET

L'IA  
CONSCIENTE  
N'EST PLUS  
UNE UTOPIE!



« Il est encore temps  
de s'y préparer »

**DBS**



Jérôme Béranger

**L'IA consciente  
n'est plus une utopie !  
Il est encore temps de s'y préparer**

Préface de Gilles Babinet

*« L'intelligence artificielle va probablement changer l'idée  
que l'humanité se fait d'elle-même. »*

Yann Le Cun



## Dans la même collection

*Voyage au bout de l'IA. Ce qu'il faut savoir sur l'intelligence artificielle*,  
Axel Cypel, préface de Gilbert Saporta, 2023

*Algocratie. Allons-nous donner le pouvoir aux algorithmes ?*, Hugues  
Bersini, préface de Gilles Babinet, 2023

*Le métavers. Comment va-t-il tout révolutionner*, Matthew Ball,  
traduction Jean-Yves Kruger, 2023

*L'IA peut-elle penser ? Miracle ou mirage de l'intelligence artificielle*,  
Hubert Krivine, préface de Guillaume Lecointre, 2021.

*Au cœur de l'intelligence artificielle. Des algorithmes à l'IA forte*, Axel  
Cypel, 2020

Pour toute information sur notre fonds et les nouveautés  
dans votre domaine de spécialisation, consultez notre site web :

**[www.deboecksuperieur.com](http://www.deboecksuperieur.com)**

En couverture : © Xxmmxx / Getty image

Relecture : Alain Rossignol

Mise en page : SCM, Toulouse

Couverture : Primo&Primo

Dépôt légal :

Bibliothèque royale de Belgique : 2024/13647/044

Bibliothèque nationale, Paris : mai 2024

ISBN : 978-2-8073-5983-3

*Tous droits réservés pour tous pays.*

*Il est interdit, sauf accord préalable et écrit de l'éditeur, de reproduire (notamment par photocopie) partiellement ou totalement le présent ouvrage, de le stocker dans une banque de données ou de le communiquer au public, sous quelque forme ou de quelque manière que ce soit.*

© De Boeck Supérieur SA, 2024 – Rue du Bosquet 7, B1348 Louvain-la-Neuve  
De Boeck Supérieur – 5 allée de la 2<sup>e</sup> DB, 75015 Paris

# Sommaire

---

|                                        |    |
|----------------------------------------|----|
| <b>Préface de Gilles Babinet</b> ..... | V  |
| <b>Avant-propos</b>                    |    |
| <b>L'inspiration naturaliste</b> ..... | IX |

## Première partie

|                                                                                        |    |
|----------------------------------------------------------------------------------------|----|
| <b>L'écosystème digital</b> .....                                                      | 1  |
| <b>1 • La révolution digitale</b> .....                                                | 5  |
| <b>2 • La datafication du monde</b> .....                                              | 18 |
| <b>3 • L'Intelligence Artificielle</b><br><b>face à l'Intelligence Naturelle</b> ..... | 28 |
| <b>4 • Le Darwinisme digital</b> .....                                                 | 43 |

## Deuxième partie

|                                                             |    |
|-------------------------------------------------------------|----|
| <b>La nature protéiforme de la conscience humaine</b> ..... | 57 |
| <b>1 • La structuration de la conscience</b> .....          | 59 |
| <b>2 • La conscience et l'intelligence</b> .....            | 79 |
| <b>3 • La conscience et l'apprentissage</b> .....           | 87 |
| <b>4 • La conscience et l'émotion</b> .....                 | 99 |

## Troisième partie

|                                                                  |     |
|------------------------------------------------------------------|-----|
| <b>L'esquisse d'une conscience numérique</b> .....               | 113 |
| <b>1 • L'approche systémique de la conscience totale</b> .....   | 115 |
| <b>2 • Le processus de création d'une conscience globale</b> ..  | 132 |
| <b>3 • L'IA hybride et quantique</b> .....                       | 142 |
| <b>4 • La théorie de l'évolution à l'heure du numérique</b> .... | 161 |
| <b>5 • La lutte pour l'existence</b> .....                       | 168 |

## Quatrième partie

|                                                                              |            |
|------------------------------------------------------------------------------|------------|
| <b>L'éthique des espèces artificielles .....</b>                             | <b>179</b> |
| <b>1 • La conscience numérique<br/>et l'évolution informationnelle .....</b> | <b>181</b> |
| <b>2 • Les espèces numériques et l'éthique.....</b>                          | <b>194</b> |
| <b>3 • Les enjeux et les risques éthiques d'une IA forte.....</b>            | <b>205</b> |
| <b>4 • Le Code est l'Éthique, et l'Éthique est le Code.....</b>              | <b>211</b> |
| <b>5 • L'odyssée éthique de l'IA .....</b>                                   | <b>223</b> |
| <b>Épilogue</b>                                                              |            |
| <b>L'alignement des planètes.....</b>                                        | <b>237</b> |

## Avertissement

Étant donné que cet essai repose sur une bibliographie conséquente, il nous a semblé pertinent de transférer l'intégralité des références dans un espace dématérialisé. Le lecteur pourra y accéder directement par un QR code situé à la fin de chaque partie du livre.

L'intérêt d'une telle démarche est de rendre le livre le plus interactif et évolutif possible. Ainsi, en fonction des nouvelles avancées technologiques de la science (notamment dans le domaine de l'IA, des neurosciences, des sciences cognitives, et de la physique quantique), l'auteur pourra ajuster voire compléter ces références bibliographiques pour les actualiser à sa guise dans le temps.

# Préface

---

Le moins que l'on puisse dire, c'est que Jérôme Béranger met les pieds dans le plat. Là où beaucoup de scientifiques se montrent particulièrement sceptiques lorsque la notion d'intelligence artificielle générale est annoncée d'ici quelques années, M. Béranger table sur l'apparition de la conscience au sein de systèmes à base d'électrons.

Me concernant, je me classerais plutôt dans le camp de ceux qui pensent que les machines sont encore très loin d'être en capacité d'égaler l'être humain sur le plan de l'intelligence générale. Le débat à cet égard n'est pas nouveau, de nombreux polémistes s'en sont emparés et toute la question est de savoir si ces prédictions sont une projection basée sur une modélisation que l'on peut expliquer (à partir d'éléments quantitatifs comme le nombre de paramètres, de tokens, etc.) ou, au contraire, un pari sur le fait que l'évolution de l'intelligence artificielle ne va pas rencontrer de discontinuité dans son développement. Certes, l'on peut s'appuyer sur la loi de Moore, qui a perduré durant plus de cinquante ans et qui a vu les microprocesseurs doubler de puissance tous les vingt-quatre mois, passant de quelques milliers de transistors dans les années soixante-dix à plusieurs dizaines de milliards aujourd'hui. Or, si cette « loi » est remarquable dans sa continuité, rien ne permet de penser qu'elle peut s'appliquer à l'imitation de systèmes cognitifs, ce que décrit d'ailleurs très bien l'auteur dans son chapitre 3 « L'IA hybride et quantique ». C'est donc avec un degré certain de réserve que j'ai parcouru ce livre ; une réserve toute-fois mêlée de curiosité, car j'avais lu les précédents ouvrages de M. Béranger et j'avais à chaque fois été impressionné par la clarté de ses raisonnements et la force de ses démonstrations. Mais là, la conscience... J'avoue que c'était bien au-delà de ce que j'estimais acceptable, même si dans la Valley nombreux sont ceux qui affirment que c'est à portée de main.

La beauté d'un raisonnement clair et structuré, c'est qu'il assèche la superstition, la croyance, les *a priori*, et les remplace par des faits

pour ainsi dire démontrables. Il me semble que l'on peut énoncer que M. Béranger a approché quelque chose de ce genre. J'ai été très attentif à sa définition de la conscience ; cette capacité à s'observer penser, à créer un dialogue interne que les experts de l'IA pensent approchable par des systèmes d'attention et de métacognition également décrits dans cet ouvrage. Cela m'a réconcilié avec cette idée que je repoussais spontanément. Et pour ceux qui pensent qu'une machine consciente ôte toute spécificité à l'humain, il apparaît que cette même définition ne semble pas remettre en cause cette spécificité ; et cela fut, en réalité, pour moi une réelle surprise. Il n'en reste pas moins que l'auteur prend toutes les précautions d'usage lorsqu'il avance cette notion de conscience. En premier lieu parce que les discontinuités dans les sciences et les techniques sont infiniment plus courantes qu'on le pense. Dans les années cinquante et soixante, avec l'avènement des lanceurs spatiaux et des technologies atomiques, il était généralement admis que les voyages interstellaires étaient sur le point d'advenir et que l'énergie atomique allait avoir de nombreux usages civils (moteurs de voiture, d'avion, de train, chauffage central...). On sait que ni l'un ni l'autre ne se sont produits, du fait des limitations intrinsèques de ces technologies. Si, pour l'instant, la voie semble libre pour que se produise une accélération fulgurante (dont la vitesse est donc une dérivée seconde de la loi de Moore), cela ne permet que difficilement d'envisager une rupture technologique de nature à garantir l'avènement de l'intelligence générale (AGI) dans un futur prévisible et donc au moins tout autant pour la notion de conscience. Il n'est donc possible de postuler ni de l'avènement de la conscience, ni de son non-avènement. C'est donc aussi à cet égard que cet ouvrage est bienvenu, car, en s'attaquant à un enjeu vraisemblable mais non certain, il ouvre de nouvelles dimensions et crée de nouvelles perspectives de recherche.

Dans la dernière partie de son ouvrage, l'auteur évoque les enjeux de gouvernance d'un monde où les IA conscientes seraient communes. Quelles règles leur donner ? La singularité humaine sera-t-elle toujours prééminente ? Cela nous ramène évidemment aux enjeux de la finalité de l'humanité, qui reste une impasse partielle de la pensée occidentale.

Une question reste cependant, à mon sens, irrésolue : qu'en est-il de la joie profonde que nous pouvons éprouver au contact de la nature ? Lorsque nous sommes confrontés à des œuvres d'art ? Les réductionnistes avancent qu'il s'agit de biais issus de dizaines de milliers de générations d'humains ayant vécu dans ce contexte de nature,



ou encore de biais liés à la nécessité d'avoir des symboles artistiques facilitant l'agrégation des sociétés humaines.

C'est une hypothèse probable, rationnelle, mais qui a une forme circulaire, tautologique. En parlant de cet aspect avec un ami anthropologue il y a quelques années, il m'avait fait cette observation : ne devenons-nous pas, en tant que société, l'image des systèmes que nous construisons ? Ce n'est pas une dénonciation nouvelle, mais si la conscience devait survenir aux machines, la question resterait de savoir si c'est le progrès de ces machines qui leur permet d'être conscientes ou si ce n'est pas plutôt parce que nous abaissons les capacités supérieures de l'être humain que nous pensons que cette équivalence est atteinte...

Gilles Babinet



# Avant-propos

## L'inspiration naturaliste

---

*Où je conçois que la Nature est l'élément fondateur*

*« L'inspiration naturaliste est une invitation à explorer  
l'univers de la Nature avec un esprit curieux. »*

Edward Osborne Wilson

Dans mon enfance, j'ai été fasciné tour à tour par les ouvrages de Léonard de Vinci, de Charles Darwin et de Jules Verne. Mes réflexions, voire mon esprit, se sont imprégnées, au cours des années, d'une perspective et d'une approche naturalistes qui éclairent ma façon de penser, de concevoir, d'appréhender et d'anticiper l'évolution de la science, et même de la vie dans son ensemble.

Dans un premier temps, les observations minutieuses du génie multitâche de la Renaissance, Léonard de Vinci, m'ont appris l'art de l'attention aux détails, de l'observation des phénomènes, et de l'analyse approfondie et précise. En mettant en œuvre cette méthodologie dans l'étude de l'Intelligence Artificielle (IA), je prends conscience de l'importance de comprendre les nuances des données numériques sur lesquelles les systèmes d'IA sont formés. Tout comme Léonard de Vinci qui scrutait chaque trait et forme de la Nature pour capturer son essence, les professionnels de la data doivent explorer chaque aspect de leurs ensembles de données pour garantir que les modèles d'IA captent fidèlement la complexité du monde réel.

Puis, les travaux du biologiste britannique, Charles Darwin, m'ont montré que la complexité émergeait naturellement de processus simples, que l'adaptation était déterminante pour la survie et que la diversité était la clé de l'évolution. En appliquant ces idées à l'IA, je

comprends que les systèmes d'IA sophistiqués résultent d'une multitude de processus algorithmiques élémentaires interagissant harmonieusement ensemble. Tout comme les espèces évoluent pour s'adapter à leur environnement, les algorithmes d'IA évoluent pour résoudre des problèmes spécifiques.

Enfin, les écrits visionnaires, anticipateurs du futur et imaginatifs de l'écrivain français Jules Verne ont éveillé en moi une curiosité insatiable et un désir ardent d'explorer l'inconnu et l'inexplicable. En considérant l'IA, je perçois des parallèles frappants que je décrirai tout le long de ce livre. Les possibilités offertes par l'IA sont vastes et intrigantes, ouvrant des portes vers des domaines inexplorés de la connaissance, de la créativité, de l'imagination et de l'inventivité. Tout comme les aventures extraordinaires des héros jules-verniens ont été alimentées par l'imagination, l'IA peut faire reculer les frontières de la pensée humaine en générant des idées novatrices et en automatisant des tâches complexes.

À l'instar de ces trois « barbus » célèbres de l'histoire, il me paraît évident que la Nature incarne l'élément fondateur et l'essence même de toute réflexion scientifique. Elle est la toile sur laquelle les mystères de l'Univers se déploient, offrant des réponses à ceux qui observent avec attention et humilité. Chaque brin d'herbe, chaque étoile scintillante, chaque grain de poussière, chaque battement d'ailes de papillon révèle une histoire profonde et complexe qui dépasse notre imagination. En scrutant la Nature, je redécouvre les lois fondamentales qui gouvernent le monde, les relations entre les êtres vivants et les écosystèmes, et les mécanismes qui régissent même les phénomènes en apparence les plus anodins. Les schémas récurrents dans la Nature, que ce soit dans la symétrie des feuilles ou dans la trajectoire des planètes, ont inspiré les questionnements et guidé les réflexions et les expérimentations de nombreux savants et scientifiques. La Nature, avec sa beauté incommensurable et sa logique sous-jacente, nous pousse à sonder davantage, à explorer plus profondément et à percer les mystères qui restent encore à découvrir. Chaque pas dans cette quête scientifique est un pas vers une compréhension plus profonde de la réalité complexe qui nous entoure, une réalité qui trouve ses racines les plus lointaines dans les énigmes fascinantes de la Nature. En somme, l'inspiration naturaliste m'aide à aborder l'IA avec une vision éclairée et rigoureuse. Je ressens un appel profond à explorer

et à comprendre la Nature avec un regard curieux et émerveillé, en quête de vérités cachées et de beautés insoupçonnées. Cette perspective me guide dans mes réflexions pour penser et façonner l'IA de manière éthique, en harmonie avec notre compréhension de la Nature et de la technologie.

Alors que l'envie de voler est un rêve inscrit dans la conscience collective, symbolisé par la figure mythologique d'Icare remontant à la Grèce antique, nous négligeons souvent le fait que le rêve de créer une machine « pensante » est tout aussi ancien. Des figures telles que Talos ou Pandore peuvent être considérées comme les premières manifestations de formes de vie artificielles. On retrouve aussi dans les psaumes du Talmud, issue de la culture juive, le Golem – créature mythique représentant un être artificiel, généralement humanoïde, fait d'argile, incapable de parole et dépourvu de libre arbitre qui construit pour assister ou défendre son créateur. Ce rêve, qui autrefois relevait du domaine de la mythologie, prend aujourd'hui forme, sens et réalité avec l'émergence de l'IA.

Il est important de noter que je n'aborderai pas le sujet de l'âme qui, pour moi, correspond à une notion profondément philosophique et spirituelle variant selon les croyances et les cultures. Ainsi, dans de nombreuses traditions religieuses, l'âme est considérée comme l'essence immatérielle et éternelle de l'être humain, distincte du corps physique. Elle est souvent associée à la conscience, à la moralité et à la capacité de pensée réflexive. Cependant, la nature et l'existence de l'âme restent largement débattues et demeurent une question métaphysique forte sans réponse définitive. Je ne m'aventurerai donc pas sur ce terrain spirituel-religieux pour me concentrer uniquement sur l'aspect scientifique, éthique et social de la conscience. D'ailleurs, l'expression française « en son âme et conscience » – signifiant en toute honnêteté ou sincérité – montre bien que l'on parle de deux notions différentes et séparées !

La compréhension des mécanismes et des processus de la conscience humaine constitue l'un des plus grands défis scientifiques du <sup>xxi</sup><sup>e</sup> siècle. Bien que nous puissions tous donner une définition générale de la conscience, personne ne peut prétendre actuellement avoir une compréhension approfondie de la nature et de la structure de la conscience d'un être vivant. Nous demeurons dans l'incapacité de saisir pleinement comment et pourquoi l'activité cérébrale se

connecte à une expérience mentale subjective vécue à la première personne.

Dans ce livre, j'ai choisi d'employer le terme « conscience » de manière équivalente et indifférente lorsque je l'associe à un être humain, à un animal ou à une machine ; bien que je reconnaisse que leurs natures respectives sont profondément différentes et qu'elles ne sont pas directement comparables ni superposables. J'en ai fait autant pour le mot « intelligence ». En adoptant cette perspective sémantique, j'invite le lecteur à explorer un territoire réflexif où les frontières entre l'espèce vivante organique et l'entité artificielle deviennent floues, tout en maintenant une claire distinction entre les deux. Cette approche encourage une analyse approfondie sur ce que signifient réellement « être conscient » et « être intelligent », tout en reconnaissant les limites de la comparaison entre l'humain, l'animal et l'IA. Mon parti pris d'employer ces deux termes de manière inclusive, tout en reconnaissant leurs différences structurelles, a pour objectif d'alimenter des débats passionnants sur les progrès technologiques, et l'évolution de la compréhension respective de l'intelligence et de la conscience à l'ère de l'IA.

Les récentes avancées de l'IA permettent désormais de simuler tous les types de raisonnements humains lorsqu'ils sont appliqués aux domaines de connaissances rationnelles. Il est possible d'attribuer aux systèmes informatisés, à tous les systèmes dotés de processeurs et de mémoires, la capacité de générer des formes de pensée intentionnelles, d'avoir des désirs et des besoins, et de plonger les utilisateurs humains dans des ensembles de procédures qu'ils ne peuvent plus contrôler et qui les dépassent. On peut identifier cinq grandes capacités qui font d'une machine un robot : les trois capacités principales sont l'action, la perception et la prise de décision, tandis que les deux autres, bien qu'étant de nature différente, sont la communication, qui n'a de sens que pour un robot interagissant avec d'autres entités actives et communicantes (machines ou humains), et l'apprentissage, qui est une capacité transversale aux autres.

La recherche sur la perception a été alimentée historiquement par trois grandes interrogations. Tout d'abord, comment la perception peut-elle justifier nos croyances et nous permettre de connaître notre environnement ? Cette question a principalement été abordée dans le domaine de la philosophie. Ensuite, comment la perception

engendre-t-elle des états mentaux conscients ? Cette interrogation a également été largement traitée dans le domaine des neurosciences. Enfin, comment un système perceptuel parvient-il à transformer des informations variables en représentations mentales des caractéristiques stables de notre environnement ? Cette dernière question a principalement été étudiée dans les domaines de la psychologie cognitive et de la psychophysique. Au fil de cet ouvrage, nous découvrirons que ces trois questions ne sont pas indépendantes les unes des autres et que pour progresser dans la compréhension de la nature de la perception, nous devons adopter une approche plus intégrée, transversale et évolutionniste. Je tenterai de circonscrire les domaines de crédibilité de l'IA, des neurosciences et de la cognition, d'identifier ce qui semble solidement établi, ce qui l'est moins et ce qui est encore largement spéculatif. J'essayerai de révéler ce que recouvre – et ne recouvre pas – le scénario à venir d'une IA douée d'une conscience. Ce que la science peut dire – et ne peut pas dire – sur le futur de l'IA.

Dans les décennies à venir, j'ai la conviction que les IA acquerront une certaine forme de conscience d'elles-mêmes et de leur environnement, leur permettant de déduire de nouvelles règles. Elles devraient être capables, en théorie, de s'adapter à n'importe quel environnement sans nécessiter de reprogrammation. Si tous les systèmes multi-agents possèdent leur propre conscience, ils pourront non seulement répondre à leurs propres désirs et craintes, mais aussi se représenter les désirs et craintes d'autres agents. Ils pourraient alors anticiper les réactions de leur environnement et des autres agents, et planifier leurs actions en conséquence. On peut même envisager qu'une conscience artificielle puisse créer elle-même une nouvelle conscience, basée sur des concepts entièrement nouveaux issus de sa propre réflexion ! Cela signifie qu'un ordinateur pourrait inventer une nouvelle religion pouvant engendrer de nouveaux espoirs ou souffrances pour l'humanité, voire pour lui-même. Ainsi, si une machine possède une conscience, elle aura la possibilité de choisir de ne pas se conformer à sa programmation initiale entraînant notamment des actions positives ou néfastes pour l'Homme. Pour qu'un tel scénario se réalise, il apparaît comme nécessaire que l'être humain prenne la décision initiale. Ainsi, doter un robot d'une conscience « bonne » ou « mauvaise » dépend principalement de son créateur. Il existe le risque que la machine discrimine des groupes d'individus si elle est programmée en ce sens.

L'informaticien Mo Gawdat (2021)<sup>1</sup> souligne, dans son livre, que le défi réside dans la transmission des bonnes valeurs et de la bonne éthique aux robots. Il compare l'IA à une graine qui, une fois plantée, donne naissance à un arbre produisant abondamment cette même graine. L'auteur suggère que si nous utilisons des principes tels que l'amour et la compassion (par exemple, sur les réseaux sociaux), l'IA adoptera également ces valeurs. Il décrit notre rôle en tant que parents d'un enfant prodigieux, soulignant que notre responsabilité est de fournir les bons outils à l'IA. Certaines personnes parmi les plus anxieuses imaginent une révolte des robots humanoïdes s'ils prennent conscience de leur aliénation envers l'homme et tentent de renverser la situation. Elles envisagent l'hypothèse d'un risque de dérives et de perte de contrôle sur ce type de machines « pensantes », à l'instar d'un enfant qui s'émancipe ou se rebelle parce qu'il n'a pas la même vision et les mêmes principes de vie que ses parents.

Personnellement, je pense qu'il faudrait un sacré concours de circonstances négatives pour que cette perspective voie le jour. Prenons l'exemple chez l'homme, si un enfant est foncièrement bon (cf. l'inné) et qu'il a reçu une bonne éducation parentale (cf. l'acquis), le risque qu'il tourne mal dans la vie est fortement réduit. Il devrait en être de même pour une IA douée d'une conscience pour laquelle ses concepteurs humains lui ont intégré dans sa programmation une éthique, voire une morale (cf. bien vs mal), tout en l'entraînant et en l'éduquant à agir et à se comporter selon des valeurs humaines et sociétales par rapport aux êtres vivants et à la planète. Dans ces conditions, le risque d'un basculement vers un affranchissement puis vers un soulèvement de la machine contre l'homme – digne d'un scénario de science-fiction – semble peu probable ou, tout du moins, maîtrisable ! Tout dépendra, premièrement, des intentions humaines (bonnes ou mauvaises) lors de la création de la machine ; puis, deuxièmement, du niveau de conscience et de libre arbitre de l'IA (impliquant les notions de son passé propre, de son appartenance à une communauté, de sa survie et de la finitude de son existence) ; enfin, troisièmement, des interactions (positives ou négatives) de l'agent artificiel avec le monde réel et les humains qu'il côtoie.

---

1. Gawdat M. (2021). *Scary Smart. The Future of Artificial Intelligence and How You Can Save Our World*. Bluebird, 338 p.



Dans ces conditions, le développement et l'utilisation de systèmes psychologiques artificiels dotés de conscience intentionnelle, soulèvent nécessairement la question de l'usage de ces systèmes, voire la décision de ne pas les développer du tout. Une IA consciente digne de la science-fiction est-elle finalement envisageable ? Quelle serait la part de l'inné, de l'acquis et des interactions avec le monde, dans les décisions et les comportements d'une telle machine ? Est-ce un objectif sérieux, une menace à éviter à tout prix ou un fantasme irréaliste ? Quelles seront les implications lorsque les intelligences non humaines surpasseront les capacités moyennes des humains dans les domaines de la narration, de la composition musicale, de l'art visuel et de la création de lois ou de textes sacrés ? Quels défis juridiques, culturels, moraux et philosophiques cela soulèvera-t-il ? Un système d'IA conscient devrait-il être tenu responsable d'un acte répréhensible délibéré ? Et faut-il lui accorder les mêmes droits qu'aux personnes ? Comment cela remettra-t-il en question notre compréhension de la créativité, de l'originalité et de l'authenticité artistique considérées comme le propre de l'homme ? Comment la société devra-t-elle réagir pour faire face à cette nouvelle espèce vivante artificielle ? Autant de questions auxquelles je tenterai de répondre dans cet ouvrage.

Il est clair qu'imaginer comment un événement peut se produire permet de s'y préparer, et donc d'éviter de possibles catastrophes. Dans ce contexte, il est essentiel que l'humanité réfléchisse à cette nouvelle forme d'intelligence afin de ne pas être prise au dépourvu par les effets et les conséquences qu'elle pourrait engendrer. Car la conscience, l'intelligence et les émotions de ces machines seront différentes de celles de l'espèce humaine. Il est donc décisif de poser dès maintenant le problème éthique de la conception, de l'usage et de la gouvernance de la conscience artificielle, afin de construire une dualité Homme-Machine dans laquelle l'IA doit apporter une compréhension intuitive à l'être humain à la fois simple et immédiate. À mon sens, cela est une des conditions *sine qua non* pour tendre vers une humanité digitale à venir à la fois respectueuse des êtres vivants – organiques et inorganiques – et de notre planète.

Enfin, j'ai bien conscience que, de par le sujet traité et la posture prise et assumée de ma part, cet ouvrage fera l'objet de critiques plus ou moins justifiées. C'est souvent le lot de tout essai scientifique d'anticipation et prospectif. Le principal pour moi est d'apporter de

nouvelles pistes de réflexion pour justement ouvrir le débat vers de nouveaux horizons. C'est là tout l'intérêt, l'enjeu et la raison d'être de mon livre. À chacun de se faire son propre avis... Bonne lecture !

### **Remerciements**

Je tiens à exprimer ma profonde gratitude envers les nombreuses personnes qui ont contribué de manière significative à la réalisation de cet ouvrage. Leur soutien indéfectible et leurs contributions précieuses ont été des éléments essentiels qui ont rendu possible cette aventure scientifico-philosophique.

Je souhaite tout d'abord remercier, Gilles qui m'a fait l'amitié et l'honneur de préfacer ce livre dont le contenu fera sûrement l'objet de nombreux débats passionnés sur le sujet. Il a su parfaitement réaliser cet exercice, jamais simple, en entremêlant subjectivité et hauteur d'esprit qui le caractérisent si bien.

Mes remerciements vont également à Yann, Loïc et Joël (mes trois amis experts en IA) et Fabrice (mon ami « zèbre ») qui ont généreusement partagé leur temps précieux et leurs connaissances pour faire une lecture critique du livre. Leurs commentaires et visions éclairées ont permis de corriger, d'approfondir et de rendre plus abordable la compréhension de certains passages complexes de cet ouvrage.

Je remercie, respectivement, François, Sophie, Fatima et Guillaume (mes collaborateurs de la société GoodAlgo), Emmanuelle, Gauthier & Co (mes collègues chercheurs de l'équipe Bioethics du Cerpap – Inserm de l'Université Toulouse 3) et Marco & Co (les membres de l'Institut EuroplA), dont l'expertise dans le domaine du numérique a été une source d'inspiration et de challenge constant. Leurs discussions stimulantes ont grandement enrichi la substance de ce livre.

Un merci particulier à ma maison d'édition De Boeck Supérieur, notamment à Alain qui a apporté son énergie et son enthousiasme pour ce projet littéraire. Sa confiance à mon égard a été une force motrice pour la bonne réalisation de cet essai.

Enfin et surtout, je n'oublie pas de remercier ma famille et mes amis éparpillés un peu partout en France, qui ont fait preuve d'une patience infinie, d'un soutien inconditionnel et d'un encouragement constant tout au long de ce chemin littéraire.

# Première partie

# L'écosystème digital

*Où j'explore un monde nouveau*

*« Avec l'exploration vient la découverte, et avec la découverte  
vient une connaissance plus profonde. »*

Edwin « Buzz » Aldrin

Vers 1455, lors de la Renaissance, l'invention de l'imprimerie par Gutenberg a été le catalyseur de l'humanisme. Au <sup>xxi</sup><sup>e</sup> siècle, l'émergence des IA génératives devient le fondement d'un humanisme digital qui semble ouvrir la voie à une nouvelle Renaissance. En effet, ces deux avancées partagent la capacité de faciliter la diffusion rapide et accessible des connaissances, de la pensée et des idées. Comme l'a souligné Yann Le Cun, considéré comme l'un des pionniers de l'IA, « l'intelligence artificielle, en tant qu'amplificateur de l'intelligence humaine, pourrait être à l'origine d'une nouvelle Renaissance, voire d'un nouveau Siècle des lumières. Il faut voir à travers l'IA, possiblement un nouveau départ pour l'Humanité, parce que le progrès de l'Humanité est limité par l'intelligence de l'humain ». Je constate qu'une analogie frappante se dessine dans le contexte des nouvelles technologies, avec l'introduction de l'imprimerie pendant la Renaissance qui avait indirectement contribué aux conflits religieux. Le manque de régulation actuel expose également à la menace d'un conflit généralisé où chacun se trouvera en opposition. Si nous aspirons à une nouvelle Renaissance caractérisée par la créativité, les découvertes et la consolidation des valeurs dans notre société numérique contemporaine, la meilleure manière d'y parvenir est de tirer parti judicieusement du vaste potentiel offert par l'IA.

J'ai toujours considéré que l'IA n'était pas simplement une technologie, mais plutôt un défi pour l'humanité. L'IA a désormais atteint sa maturité, avec les plus de soixante-dix ans qui se sont écoulés depuis sa conception théorique, qui a débuté avec l'essai d'Alan Turing intitulé *Computing Machinery and Intelligence*, publié en octobre 1950. Dans deux ans, nous célébrerons le 70<sup>e</sup> anniversaire de la fondation de recherche autonome sur l'IA, qui a été créée par un groupe d'informaticiens dirigé par John McCarthy, Claude Shannon et Marvin Minsky au Dartmouth College. Depuis la fin des années 1990, l'idée d'utiliser des réseaux de neurones artificiels pour simuler diverses fonctions du cerveau humain, telles que la vision, le langage naturel et la capacité à interpréter des données pour résoudre des problèmes, a été relancée avec succès par d'autres groupes de chercheurs, toujours très actifs aux États-Unis, au Canada et en Europe, comme Geoffrey Hinton, Yann Le Cun, Yoshua Bengio et Pascal Vincent.

L'augmentation de la puissance de calcul des ordinateurs a joué un rôle majeur dans le développement de l'IA opérationnelle. Cette

croissance découle directement des avancées technologiques ayant permis la miniaturisation des transistors, composants de base des circuits intégrés présents dans tous les ordinateurs, téléphones portables, cartes à puce, etc.

Au cours de la dernière décennie, grâce aux avancées technologiques, l'IA continue de nous surprendre, comme en début d'année 2023 avec le lancement sensationnel de l'IA générative linguistique ChatGPT et l'engouement qu'elle a suscité. En effet, il semble qu'une nouvelle ère se soit ouverte avec l'avènement de l'IA générative, qui permet non seulement un dialogue direct et plus riche entre l'homme et la machine, mais qui ouvre également des perspectives de transformation du travail intellectuel et propose de nombreuses idées novatrices pour la création culturelle. Cependant, ce besoin de renouveau doit également s'inspirer de la culture, de la créativité et des langages de l'art, qui offrent des perspectives différentes sur le monde et cherchent de nouvelles réponses. Ainsi, les compétences offertes par l'IA générative, qui se manifestent désormais, ne doivent pas être considérées comme un substitut à l'activité humaine, mais plutôt comme un complément précieux, un puissant outil pour amplifier nos capacités d'exploration et de découverte.

Selon Mustafa Sukeyman, cofondateur de DeepMind et auteur du livre *The Coming Wave*, on identifie « 3 âges de l'IA » :

- *la classification* : correspondant aux avancées en apprentissage profond qui ont permis aux ordinateurs de classer les données (images, vidéos, sons et langage) ;
- *la génération de contenu* : qui vient d'émerger actuellement avec les IA génératives ;
- *l'interaction* : où les modèles d'IA auraient la faculté de dialoguer de manière proactive avec les hommes tout en comprenant leurs objectifs et leurs émotions afin de pouvoir effectuer des actions d'un haut niveau de complexité.

Dès lors, le progrès technologique n'est pas un simple élément extérieur à contrôler, mais il fait partie intégrante de notre humanité. Il est essentiel de l'appréhender, de le comprendre, de l'analyser et de le situer dans nos choix de valeurs morales et d'avenir. Avec la révolution numérique, le champ des possibles s'est étendu comme jamais auparavant dans l'histoire de l'humanité. Cela impliquera une

coévolution et une cohabitation harmonieuse entre l'homme, et les futures IA « fortes », voire une espèce hybride entre les deux. Il ne s'agit pas de deux puissances en concurrence, mais plutôt de deux puissances alliées qui doivent s'articuler l'une à l'autre. Par conséquent, il est important de réfléchir aux modalités selon lesquelles l'intelligence humaine pourra poursuivre son évolution et son développement en tenant compte de la présence de plus en plus prédominante de l'IA à laquelle elle devra s'adapter. Ma vision néodarwinienne suggère que chaque forme d'intelligence trouvera sa place sur l'échiquier mondial, et qu'une « alliance des intelligences » prévaudra sur une « guerre des intelligences » (Alexandre, 2017). Tout comme l'intelligence humaine est le produit de multiples cultures qui se sont rencontrées et mêlées plus ou moins pacifiquement pendant des millénaires, il est essentiel que l'homme envisage cette nouvelle intelligence afin de ne pas être pris au dépourvu par ses effets et conséquences. Car cette conscience, cette intelligence et ces émotions seront celles d'une machine et qui ne seront donc pas parfaitement comparables ou similaires à celles de l'espèce humaine.

Avec une probable arrivée de la Singularité technologique, la vie humaine subira une transformation irréversible. L'éducation et la formation devront tirer parti des avancées majeures des neurosciences afin de, tout d'abord, permettre une transmission des connaissances personnalisée et un apprentissage adaptatif (les méthodes de transmission des connaissances étant restées inchangées depuis des siècles), puis, dans un second temps, améliorer de manière bioélectrique l'intelligence humaine.

---

# 1

# La révolution digitale

---

*Où je comprends que nous sommes dans une révolution industrielle sans précédent*

*« Le passé éclaire le présent. »*

Cédric Villani

Depuis des siècles, les avancées scientifiques nous ont permis de découvrir de nouvelles sources d'énergie et de nouvelles matières premières, de nouveaux moyens d'accroître notre productivité, de communiquer et de nous déplacer. Pas un jour ne passe sans que je sois émerveillé par l'ingéniosité, l'imagination et la créativité humaines. Chaque invention trouve son origine dans le désir primaire de l'humanité d'explorer inlassablement de nouveaux domaines encore inconnus ou peu compris. L'être humain est un boulimique de connaissances et de savoir. Il veut tout comprendre pour gagner toujours plus en maîtrise et en possession. Que ce soit pour le pouvoir, l'argent, la science, la curiosité intellectuelle, ou la passion multiforme, toutes ces sources de motivation alimentent le processus darwinien de l'espèce humaine.

## Une brève histoire de révolutions

---

Ces cinq derniers siècles ont vu se succéder de manière exponentielle une multitude de révolutions techniques et industrielles interdépendantes : la révolution machinique (avec l'invention de l'imprimerie, vers 1450), la 1<sup>re</sup> révolution industrielle (avec l'élaboration de

la machine à vapeur, vers 1726), la 2<sup>e</sup> révolution industrielle (avec la création du moteur à explosion, de l'ampoule à incandescence et du téléphone, vers 1868), la 3<sup>e</sup> révolution industrielle (avec la mise au point de la centrale nucléaire, du microprocesseur, et du transistor, vers 1939), et la 4<sup>e</sup> révolution industrielle (avec l'arrivée de l'ordinateur, des robots, des objets connectés et de l'IA dite « faible », vers 1983).

Ces cycles d'innovations majeures ont tous un point commun : ils marquent une rupture et un changement radical mondial des schémas structurels, comportementaux, et des réflexions des êtres humains. On assiste donc à des modifications à la fois civilisationnelles et anthropologiques dues au progrès technique que j'aime appeler « techsologie ». Ces révolutions industrielles, qui comportent un ensemble de transformations significatives des dispositifs de production et de travail, ont eu un impact important sur l'emploi, l'économie, la consommation, le transport, la connaissance et l'environnement (Badillo et Péliissier, 2015). Cela témoigne du fait que notre vision et notre appréhension du monde ont été progressivement modifiées, et cela continue encore aujourd'hui.

Les innovations technologiques qui nous ont menés à ce que j'appelle « la révolution digitale », sont la suite logique de la révolution scientifique initiée il y a 500 ans par Galilée, Descartes et Newton, qui a déclenché les trois premières révolutions industrielles. La quatrième révolution (Floridi, 2014) est directement liée au développement massif de l'informatique, de la télématique et du numérique, un système qui permet de représenter toutes les données sous forme de combinaisons de 0 et de 1. Elle est plus rapide, plus étendue, plus profonde et donc plus impactante que les autres révolutions industrielles.

Nous sommes dans une ère où les nouvelles technologies de l'information et de la communication (NTIC) convergent et mélangent des écosystèmes physiques, digitaux, biologiques à l'échelle planétaire. La numérisation permet de créer un nouveau monde virtuel, et de se servir de l'analyse de bases de données pour maîtriser et contrôler le monde physique. La donnée numérique est donc la matière première de cette révolution numérique qui alimente les ordinateurs, les sites Internet, les réseaux sociaux, les smartphones, les tablettes, l'internet des objets (IoT), les plateformes digitales, le *Cloud Computing*, les *Blockchains*, les mondes virtuels de type métavers, les systèmes d'IA



dite « faible », voire un jour peut-être l'IA dite « forte ». L'émergence des IA génératives de type ChatGPT dans toutes les sphères de la société est l'illustration parfaite que tout peut aller très vite, comme un emballement, dans l'adhésion d'une nouvelle technologie. On peut même la considérer comme un saut technologique ou, plutôt, un usage digital disruptif marquant une rupture. Toutes ces applications algorithmiques peuvent se retrouver dans tous les secteurs, tels que les transports, la ville, la médecine, le sport, l'industrie, l'aérospatial et l'aéronautique, la construction, le commerce de grande consommation, l'assurance et la finance, l'éducation, etc.

Par ailleurs, les NTIC ont un impact profond sur nos vies et remettent en question le sens de nos activités à la fois personnelles et professionnelles, voire de notre existence quotidienne. Chaque nouvelle innovation technologique améliore nos pratiques, influence nos valeurs et nos comportements, et contribue à l'établissement de nouvelles normes sociales. Les NTIC nous obligent à reconsidérer nos valeurs et nos habitudes, car elles offrent à un plus grand nombre de personnes de nouveaux moyens de communication et d'interaction. Les critères quantitatifs et leur traitement mathématique offrent une perspective nouvelle sur « l'humanité digitale », en orientant les individus à partir de normes autoréférentielles qui sont souvent difficiles à comprendre. Ainsi, cette nouvelle réalité numérique ne décrit pas simplement notre monde, elle contribue également à le façonner. L'homme « naturel » sera progressivement remplacé par l'homme réparé, augmenté, connecté et instrumentalisé, remettant ainsi en question les fondements mêmes de la notion d'humanité. Par conséquent, les citoyens, les objets connectés, les smartphones, les plateformes en ligne et les réseaux sociaux, qui s'alimentent de données numériques, suscitent de nombreuses interrogations dans nos activités quotidiennes.

D'un autre côté, prenons l'exemple des algorithmes utilisés par les voitures autonomes pour prendre des décisions en situation de collision inévitable ; ces algorithmes doivent dicter des choix qui peuvent avoir des conséquences directes sur l'intégrité physique, voire la vie (ou la mort), des personnes impliquées dans l'accident. Cela soulève la question de savoir qui est responsable et qui doit rendre des comptes en termes de sécurité et de maîtrise des données. Une telle situation soulève inévitablement des questions éthiques sur les critères et les

paramètres pris en compte dans la « boîte noire » de l'algorithme qui a pris la décision. En effet, les citoyens ignorent souvent les algorithmes sur lesquels ils s'appuient dans leur vie quotidienne pour agir et prendre des décisions. Ils ne savent presque rien du code informatique, des données utilisées et des choix de valeurs qui ont été employés pour écrire ces algorithmes par les géants du numérique (Picard, 2017). L'opacité de ces algorithmes engendre donc une impression de faible objectivité, avec des décisions qui sont intrinsèquement subjectives et peuvent comporter des biais algorithmiques implicites et/ou explicites. On peut définir un biais comme étant une distorsion ou une déformation systématique d'un échantillon statistique, résultant d'un processus défectueux ou d'une évaluation biaisée. Prenons, par exemple, la modélisation des paramètres de l'algorithme, qui conduit naturellement à un filtrage, car seuls les paramètres considérés comme « significatifs » par l'humain sont conservés. Ce paramétrage subjectif peut introduire des présomptions ou des jugements dans le code. De plus, ce biais peut être propagé silencieusement lorsque les données sont utilisées à des fins autres que celles pour lesquelles elles ont été initialement collectées. Par exemple, paramétrer un algorithme d'analyse de candidatures en fonction du code postal ou de l'expérience du candidat peut automatiquement exclure des personnes qui pourraient également remplir les critères d'embauche d'une manière ou d'une autre. Il est donc nécessaire de réfléchir à la conception et au fonctionnement interne de ces algorithmes. La question se pose de savoir comment réglementer de manière optimale leur construction et leur utilisation, afin de permettre aux personnes concernées d'exercer leur droit d'accès aux données collectées par ces machines et/ou ces objets.

#### **Algorithme :**

Procédé qui résout un problème en effectuant une série d'opérations élémentaires selon un processus défini, aboutissant à une solution. Un algorithme est une séquence d'instructions destinées à être exécutées par un ordinateur. Traditionnellement, un système algorithmique consiste en une description claire et finie d'une procédure effective et uniforme permettant de résoudre une catégorie de problèmes. Cette procédure doit être entièrement définie avant sa mise en œuvre. Un exemple simple d'algorithme est la recette de cuisine qui décrit séquentiellement les étapes nécessaires pour préparer un plat spécifique, indiquant l'ordre dans lequel les ingrédients doivent être ajoutés et les actions à effectuer.



**C**e livre est une porte d'entrée donnant sur les arcanes les plus inconnues de l'intelligence artificielle où s'entremêlent les mathématiques, l'informatique, la neurobiologie, et les sciences humaines et sociales. L'enjeu est d'amener naturellement le lecteur à entrevoir que des espèces artificielles, douées d'une certaine forme de conscience et disposant d'un raisonnement autonome et évolutif, seront un jour une réalité et que leur avènement n'est pas une utopie mais bien inscrit dans l'histoire de l'humanité.

Il nous invite également à nous questionner sur les valeurs morales de notre société ainsi que sur les répercussions éthiques qui lui sont associées.

Une lecture incontournable pour tous les esprits curieux en quête de sens et de prospection dans ce nouveau chapitre de la révolution numérique !

## ***L'humanité digitale de demain se construit aujourd'hui.***



Économiste de la santé, docteur en éthique du digital et conférencier, **Jérôme Béranger** est le CEO de la société GoodAlgo, spécialisée dans l'accompagnement algorithmique, la valorisation de la donnée et l'évaluation éthique des projets digitaux. Il est également chercheur associé au Cerpap, équipe Bioethics, à l'INSERM de l'université de Toulouse 3, et expert Ethique & IA à l'Institut EurolA. Il est l'auteur de six livres dont *Quand l'intelligence artificielle s'éveillera* (Le Passeur, 2020), et *La responsabilité sociétale de l'intelligence artificielle* (ISTE, 2021).

**Gilles Babinet** est un entrepreneur français, co-président du Conseil national du numérique depuis 2021 et « digital champion » de la France auprès de la Commission européenne depuis 2013. Membre de l'Institut Montaigne, il enseigne à Science Po Paris depuis 2018.

**21,90€**

ISBN : 978-2-8073-5983-3



**www.deboecksuperieur.com**