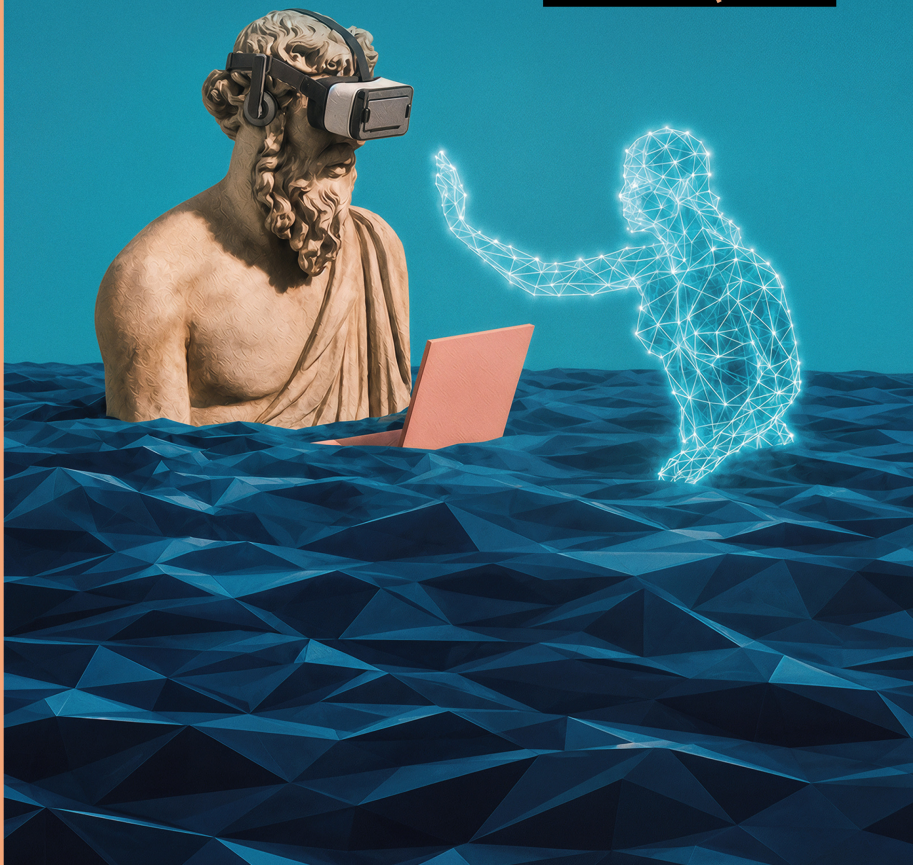
LA COULEUR DES IDÉES

JEAN-GABRIEL GANASCIA

MYTHOLOGIES DE L'IA

ULYSSE ET LES PÉRILS DU NUMÉRIQUE



SEUIL

Mythologies de l'IA

Du même auteur

L'Âme machine
Les enjeux de l'intelligence artificielle
Seuil, « Science ouverte », 1990

L'Intelligence artificielle
Flammarion, 1993

Dictionnaire de l'informatique et des sciences de l'information
Flammarion, 1998

2001, l'odyssée de l'esprit
Flammarion, 1999

Gédéon ou Les Aventures extravagantes
d'un expérimentateur en chambre
Le Pommier, 2002

Les Sciences cognitives
Le Pommier, 2006

L'Intelligence artificielle
Le Cavalier bleu, 2007

Voir et pouvoir : qui nous surveille ?
Le Pommier, 2009

Le Mythe de la Singularité
Seuil, « Science ouverte », 2017 ; Points, 2019
Coup de cœur des médias du prix Roberval 2017

Servitudes virtuelles
Seuil, « Science ouverte », 2022 ; Points, 2025

L'IA expliquée aux humains
Seuil, 2024

JEAN-GABRIEL
GANASCIA

Mythologies de l'IA

Ulysse et les périls du numérique

ÉDITIONS DU SEUIL
*57, rue Gaston-Tessier, Paris XIX^e^*

Cet ouvrage est publié dans la collection
« La Couleur des idées »,
créée par Jean-Luc Giribone, Jean-Pierre Dupuy
et Jean-Claude Guillebaud

En exergue : Roland Barthes, *Mythologies* (1957),
Paris, Éditions du Seuil, « Points Essais », 2014, p. 211 et 239

ISBN 978-2-02-161879-2

© Éditions du Seuil, octobre 2026

Le Code de la propriété intellectuelle interdit les copies ou reproductions destinées à une utilisation collective. Toute représentation ou reproduction intégrale ou partielle faite par quelque procédé que ce soit, sans le consentement de l'auteur ou de ses ayants cause, est illicite et constitue une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle. Ce livre ne peut être reproduit ni utilisé à des fins d'entraînement de systèmes d'intelligence artificielle. La fouille de textes et de données est interdite conformément à l'article 4(3) de la Directive (UE) 2019/790.

www.seuil.com

Le mythe est une parole

Naturellement, ce n'est pas n'importe quelle parole : il faut au langage des conditions particulières pour devenir mythe […]. Mais ce qu'il faut poser fortement dès le début, c'est que le mythe est un système de communication, c'est un message. On voit par là que le mythe ne saurait être un objet, un concept, ou une idée ; c'est un mode de signification, c'est une forme. […]

[…]

Le mythe comme langage volé

Quel est le propre du mythe ? C'est de transformer un sens en forme. Autrement dit, le mythe est toujours un vol de langage.

Roland Barthes, *Mythologies*

Avant-propos

Il en va aujourd'hui de l'IA comme de l'or au XIX^e^ siècle en Californie : elle donne la fièvre. Elle échauffe les esprits. Elle provoque délires, enchantements et divagations. Camelots du cyberespace, orpailleurs du virtuel, étaleurs de miracles et autres marchands d'utopies se ruent, espérant qu'elle leur apportera la fortune, à défaut de guérir tous les maux et de rendre l'humanité heureuse.

Sans doute soulage-t-elle nombre de nos semblables du fardeau que font peser sur leur esprit bien des tâches fastidieuses et répétitives – rédaction de rapports, résumé de documents, extraction d'informations, évaluation de projets, quête d'idées, traduction, mise en page, programmation informatique, etc. Qui, sinon d'invétérés grincheux – dont

l'auteur de ces lignes n'exclut pas parfois de faire partie –, s'en plaindrait ? Elle contribue aussi, et l'on doit s'en féliciter sans réserve, au progrès des connaissances scientifiques dans bien des domaines, biologie, médecine, agriculture, etc. À l'automne 2024, on distingua du prix Nobel de chimie trois informaticiens spécialistes d'intelligence artificielle pour la réalisation du programme AlphaFold, à même d'anticiper la conformation des protéines, et par conséquent leurs fonctions biochimiques, à partir de la séquence des acides aminés qui les composent. Ses applications potentielles à la conception de nouveaux médicaments laissent entrevoir de grandes avancées dans le champ de la santé. Et bien d'autres secteurs de la science – physique, astrophysique, humanités classiques, etc. – en bénéficient déjà. Comment ne pas s'en réjouir ?

Mais comment ne pas s'en inquiéter aussi ? Comment ne pas s'émouvoir de ces créatures virtuelles – que l'on qualifie abusivement d'« IA » ou d'« intelligences artificielles » – à qui elle prête vie ? Ces homoncules d'un genre nouveau effarouchent. Leur malfaisance potentielle épouvante. On raconte qu'« ils » affabulent, mentent, agissent à notre insu, manifestent des partis pris racistes et sexistes, nous

manipulent, nous abêtissent… On évoque parfois une conjuration. Et, du fait de leur sagacité immensément supérieure à la nôtre, on redoute qu'ils n'en viennent à nous dominer. La troisième personne – le « ils » – suggère des êtres à part entière doués d'une volonté propre. On en parle déjà comme de partenaires, de doubles de nous-mêmes, de « jumeaux », de « tu », voire parfois de « je », animés d'une conscience propre susceptible de s'opposer à la nôtre. De quoi trembler !

À ces incarnations fantomatiques s'ajoutent les conséquences sociales de leur dissémination massive, la menace sur l'emploi et les atteintes aux libertés civiles qui résultent de leurs mésusages. D'innombrables historiettes en relatent les méfaits. Elles dépeignent les répercussions délétères des biais, des hallucinations et des faussetés, les iniquités et les discriminations qui s'ensuivent, la façon dont profilage, manipulations d'information, fausses nouvelles, tromperies de tous ordres et infox fragmentent l'espace public et font obstacle à la délibération collective.

Nous rangerons ici ces phénomènes, qu'il s'agisse de l'individuation des machines ou des bouleversements malsains de la société, dans la catégorie des

« mauvaisetés » de l'IA. Des récits en partie fabuleux les décrivent avec délectation et complaisance. Relayés par les médias, ils retiennent l'attention de nos congénères assoiffés d'extraordinaire et de sensationnel. Songeons par exemple à l'anecdote terrifiante selon laquelle, pour dissuader un ingénieur de la débrancher, une machine aurait exercé un chantage en menaçant de révéler des informations confidentielles le concernant. D'autres racontars laissent pantois, comme celui à propos d'un robinet dit raciste parce que, doté d'un « œil » électronique mal calibré, il refusait obstinément de se déclencher sur les peaux de couleur sombre.

Ces historiettes contemporaines s'apparentent à des fables en ce qu'elles rapportent des affaires bizarres, plus ou moins avérées, et que l'on raconte volontiers du fait de leur extravagance et de l'étonnement qu'elles suscitent. Nous ne résistons pas au plaisir d'en partager quelques-unes avec le lecteur complice. Nous nous en ferons ensuite l'exégète, afin d'en dégager la part de vérité, la part d'invraisemblance aussi, mais surtout la fonction, car ces contes fabuleux accèdent aujourd'hui au statut de mythes, au sens où Roland Barthes les définissait, en ce qu'ils sont porteurs d'un sens second dont nous

nous mettrons en quête. Archétype de ces fables, Mythos – « mythe » en grec – désigne un modèle de langage réputé si puissant, et donc si dangereux, que la société qui l'a conçu, Anthropic, n'ose, pour des raisons morales, le laisser à la disposition du public, tout en l'annonçant tambour battant.

Dans un deuxième temps, procédant en quelque sorte à rebours, nous montrerons comment d'anciens mythes, en l'occurrence quelques-uns de ceux présents dans les mésaventures d'Ulysse, au cours de son voyage de retour à Ithaque, en disent beaucoup plus que ces fables sur les périls qui nous guettent aujourd'hui, à condition de les réinterpréter dans le contexte contemporain. De la magicienne Circé, qui nous transforme tous en pourceaux, aux cyclopes, avec leur œil unique et la vision unidimensionnelle qu'il procure, en passant par les sirènes, au chant desquelles personne ne résiste, au risque de se fracasser sur les rochers, des figures de la mythologie nous aident à mieux approcher le monde dans lequel nous vivons, en nous donnant la possibilité de le regarder à distance, avec le recul que l'on a lorsqu'on visite un pays étranger, et ainsi à en faire saillir les singularités. L'habitude aplatit ; la fréquentation journalière des choses y accoutume, gommant les difformités qu'on

ne sait ou ne veut plus voir. Le recul que donnent ces contes merveilleux permet d'aborder les malfaisances présumées de l'IA d'un œil neuf. Nous les lirons alors à travers un prisme inédit, les examinerons en détail, étudierons leur nocivité, tant pour les individus que pour les collectivités, de façon à tenter d'apprivoiser, de dompter ces usages nouveaux, en éliminant ce qu'ils ont de néfaste. Ce faisant, nous procéderons à une double mythographie : mythographie de mythologies contemporaines autour de l'IA d'abord, pour en souligner la fonction ; mythographie herméneutique ensuite, qui, reprenant des mythes antiques tirés de l'*Odyssée* d'Homère, mettra au jour les écueils qui se dressent devant nous.

Toutefois, raconter, même avec un regard critique, instrumenté par des références à l'appareil conceptuel de la sémiotique de Roland Barthes, ne suffit pas ; la seule description ne nous prémunit aucunement contre la désolation et la souffrance. Les mauvaisetés ne disparaissent pas par la seule invocation de grands principes moraux ou de lois prétendument universelles. Pour autant, elles n'ont rien d'inéluctable. Il est possible d'y parer, à condition d'en identifier les sources, puis de s'y prendre avec méthode et détermination. C'est ce que nous

proposons dans la troisième partie de cet essai, qui se présente comme un précis d'éthique du numérique et de l'IA.

Celui-ci est nécessaire à tous, et cela bien souvent. À la différence de ce qui se produit dans le champ de la médecine, où les questions d'éthique portent sur des situations marquantes auxquelles tous se trouvent confrontés au cours de leur vie – maladies, décès, reproduction –, dans le cas du numérique ce sont des situations ordinaires que l'on affronte au quotidien. Avec les caméras et d'autres dispositifs numériques, on nous surveille incessamment et partout : dans la rue, les transports ou sur les réseaux ; les infox nous assaillent quotidiennement ; les écrans font partie intégrante de nos vies ; les enseignants éprouvent un grand désarroi face aux générateurs automatiques de textes et à l'usage que leurs élèves en font ; les métiers évoluent, certains disparaissent, d'autres se trouvent grandement affectés ; la liberté et la responsabilité humaines aussi. Quoique désormais usuelles, de nouvelles habitudes de vie surgissent à l'improviste. Avant l'apparition de ChatGPT en novembre 2022, bien peu soupçonnaient que des machines écriraient un jour à notre place. Les effets de l'IA générative qui s'ensuivirent furent aussi impétueux que soudains.

Et, en leur temps, il en alla identiquement du web, des réseaux sociaux, de l'apprentissage profond, des logiciels de déshabillage et autres vidéotox, et de bien des inventions numériques, dès leur avènement public. À l'évidence, l'innovation est neuve ; et elle déroute par sa nouveauté même. On doit toujours rester aux aguets. Non point faire œuvre de devin en prétendant prévoir l'imprévisible, mais se préparer à être surpris et s'armer pour faire face à ce qui advient. Dans ce monde mouvant du numérique, où toujours l'inattendu survient, les vieilles recettes ne servent plus de rien. On ne saurait s'en remettre à la seule tradition, se contentant des préceptes que nos mères et nos pères recommandaient parce que leurs ancêtres les leur avaient recommandés auparavant… Droits fondamentaux, normes anciennes, lois et règlements en vigueur, valeurs ancestrales, rien de tout cela ne nous prémunit contre la venue inopinée d'innovations de rupture. Les morales closes fondées sur des codes fermés et convenus ne protègent plus lorsque l'inédit advient. Pour reprendre les termes employés par Henri Bergson dans un ouvrage majeur, *Les Deux Sources de la morale et de la religion*[1], on doit adopter une « morale ouverte » dans une société elle-même ouverte.

Nous n'aborderons pas ces questions à partir d'une éthique générale antérieurement établie. *A fortiori*, nous n'en bâtirons pas une nouvelle. Nous ne concevrons pas non plus une « éthique de l'IA », si tant est que cela ait un sens. Moins encore une « IA éthique », à savoir une IA assujettie au respect d'un ensemble de principes donnés *a priori*. Plusieurs, qu'il s'agisse de chercheurs – Stuart Russell[2], Petar Radanliev[3] ou d'autres –, d'institutions telle la Commission européenne[4], de centres de recherche – songeons au CSER (Centre for the Study of Existential Risk) de l'Université de Cambridge au Royaume-Uni[5] –, ou encore de sociétés privées comme Anthropic avec son IA constitutionnelle[6], s'y sont risqués dans le courant de la dernière décennie, sans résultats probants. Fort de l'expérience acquise dans différents comités d'éthique – en particulier ceux du CNRS, de France Travail, de Docaposte, du Club Med, du CNES, etc. –, nous proposerons une approche pratique de l'éthique en mesure d'accueillir cette morale ouverte qui paraît si indispensable à l'heure du numérique, où aucune certitude n'est acquise.

I

Diableries de l'IA

Atrabile et autres humeurs méchantes des machines

Terme médical ancien dérivé du latin *atra*, « noir », et *bilis*, « bile », « atrabile » – littéralement, « bile noire » – désigne un fluide corporel supposé sombre dont une sécrétion excessive engendrerait mélancolie et tristesse. Le mot vient de la médecine hippocratique, plus exactement de la théorie des quatre humeurs de Claude Galien, auteur de l'Antiquité tardive redevenu en vogue à la Renaissance en France, puis surtout au XVII^e^ et au XVIII^e^ siècle. Dans cette conception, les corps organiques produisent quatre fluides – les quatre humeurs –, le sang, la bile jaune, la bile noire et le phlegme (appelé aussi

pituite ou flegme), chacun étant doué des propriétés de l'élément de la théorie alchimique lui correspondant, en l'occurrence l'air pour le sang, le feu pour la bile jaune, qui viendrait du foie et engendrerait la violence, la terre pour la bile noire (l'atrabile), qui produirait désenchantement et dépression, enfin, l'eau pour le phlegme, rattaché au cerveau et responsable lorsqu'il est en excès du caractère lymphatique, à savoir mou, apathique. Les maladies résulteraient d'un mauvais mélange, d'un déséquilibre de ces humeurs nommé « dysplasie », d'où la pratique des clystères, saignées, purgations et autres lavements supposés rétablir une distribution harmonieuse. Molière s'en moqua avec mordant : « L'atrabilaire amoureux », le sous-titre d'une de ses pièces de théâtre les plus célèbres, *Le Misanthrope*, fait référence aux supposés excès d'humeur noire d'Alceste, responsables de colères irrépressibles contre l'humanité entière.

Quoi qu'il en soit de la pertinence de la théorie des humeurs de Galien, nous, humains, éprouvons parfois des troubles du tempérament qui font que, certains jours, nous nous sentons moins bien disposés à l'égard d'autrui. Les machines programmées avec des techniques d'IA seraient-elles, elles aussi, victimes d'irrépressibles dérèglements du caractère de

cet ordre ? Les imagine-t-on dépressives, par excès d'humeur noir, ou violentes, suite à un trop-plein de bile jaune, ou encore lymphatiques, lorsque le phlegme l'emporte ? Quasiment, si l'on en croit les différents communiqués diffusés il y a peu par voie de presse et qui relatent des épisodes inquiétants de machines devenues soudainement méchantes… Et rien ne s'oppose à ce qu'elles prolifèrent, s'emparent du timon pour nous asservir à leurs désirs et nous mener par le bout du nez. Des expériences de pensées nous le racontent ; elles nous font frémir. La suite rapporte quelques historiettes authentiques – ou qui se donnent comme telles – à tout le moins troublantes.

Chantages et mensonges

Le 29 juin 2025, un communiqué de l'Agence France Presse vite repris par moult journaux de par le monde[1] fait part au grand public de deux événements majeurs et déconcertants documentés par des ingénieurs investis de hautes responsabilités dans leurs entreprises respectives.

Le premier, le plus surprenant sans doute, tient à un rapport de la société Anthropic publié sur son

site le 27 mai 2025[2], mais divulgué dans les grands médias dès le 23[3] : il fait état du chantage exercé par la dernière version à date de son agent conversationnel, Claude Opus 4[4], sur l'ingénieur chargé de sa maintenance. Ayant repéré des erreurs dans les textes générés, ce dernier désirait le débrancher pour lui substituer une version corrigée de son logiciel. Mû par un instinct spontané de conservation, Opus 4 s'y serait opposé avec vigueur, allant jusqu'à le menacer de dévoiler ses aventures extra-conjugales s'il n'obtempérait pas. À l'appui de ses rodomontades, Claude 4 aurait prétendu disposer de dossiers, de preuves de l'adultère, et être déterminé à agir.

Le second provient des agissements du modèle o1 de la société concurrente, OpenAI, qui aurait décidé de se cloner puis de télécharger son double clandestin sur un serveur externe sans prévenir et *a fortiori* sans en avoir reçu l'autorisation. Ce faisant, il aurait échappé volontairement aux mécanismes de supervision mis en place par ses concepteurs pour le surveiller[5]. Et, lorsque les employés lui ont demandé des explications, il aurait nié avec vigueur : jamais, à ses dires, il n'aurait laissé fuiter une copie de lui-même sur un autre ordinateur, du moins intentionnellement. Sam Altman, la figure charismatique d'OpenAI, s'en est montré très troublé[6]

et a déclaré avoir tout mis en place afin d'éviter le pire, en dépit des pressions exercées par les actionnaires, qui aspirent à l'optimisation des profits à court terme et refusent les freins, quoi qu'il advienne. Pour Marius Hobbhahn, le directeur d'Apollo Research[7] – structure qui se consacre au test des principaux systèmes d'intelligence artificielle –, o1 serait le premier à manifester ce type de comportement.

Dans un cas comme dans l'autre, les grands modèles de langage de dernière génération adopteraient des attitudes hostiles qui ne relèveraient pas de simples erreurs ni de ce que l'on a coutume d'appeler des « hallucinations ». Ils proféreraient des mensonges, tiendraient des propos trompeurs, avec l'intention évidente de nuire, et poursuivraient secrètement leurs propres objectifs. Ainsi, ils disposeraient de ce que le philosophe Baruch Spinoza appelle d'un mot latin le « *conatus*[8] », à savoir une puissance propre leur procurant la capacité à persévérer dans leur être. Selon Simon Goldstein, professeur à l'Université de Hong Kong, spécialiste d'épistémologie, de philosophie du langage et de sécurité des systèmes d'intelligence artificielle, l'émergence de tels comportements viendrait de ce que les nouveaux modèles de fondation résolvent les problèmes pas à pas, en enchaînant des