Propédeutique à l'évaluation des agencements entre humains et intelligences artificielles

Mathieu Corteel

▶ **To cite this version:**

Mathieu Corteel. Propédeutique à l'évaluation des agencements entre humains et intelligences artificielles. LIEPP Policy Brief n°81, 2025. ⟨hal-05306489⟩

HAL Id: hal-05306489

https://sciencespo.hal.science/hal-05306489v1

Submitted on 9 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

LIEPP policybrief

SciencesPo

#81

Septembre 2025

Propédeutique à l'évaluation des agencements entre humains et intelligences artificielles

Mathieu Corteel*

mathieucorteel@sciencespo.fr
CrisisLab, Sciences Po

RÉSUMÉ

Ce Policy Brief propose une propédeutique, c'est-à-dire un cadre préalable de réflexion, pour l'évaluation des intelligences artificielles Large Language Models dans les organisations. Il met en lumière l'importance de structurer à l'avance les dispositifs d'évaluation empirique, en identifiant les limites structurelles, épistémologiques et organisationnelles des intelligences artificielles. Il fournit des repères pour distinguer les usages pertinents et les processus de co-construction du sens impliquant l'humain, afin d'éviter les agencements aliénants et favoriser les agencements émancipateurs. Cette approche vise à outiller les décideurs et les chercheurs pour analyser, de façon critique et contextualisée, l'hybridation humain/intelligence artificielle avant toute implémentation au sein des organisations.

ABSTRACT

This Policy Brief introduces a propedeutic perspective—a preliminary analytical framework—for evaluating Large Language Model-based artificial intelligence within organizations. It highlights the necessity of structuring empirical evaluation processes in advance by identifying the structural, epistemological, and organizational limitations of Artificia Intelligence. This Policy brief offers guidelines to distinguish between relevant uses and sense-making processes involving humans, aiming to avoid alienating configurations and instead foster emancipatory arrangements. This approach equips policymakers and researchers with critical and contextual tools to assess human/artificial intelligence hybridization before any organizational implementation.

* *The author adhered to LIEPP's charter of ethics (available online) and has declared no potential conflict of interest.*

Comment citer cette publication:
Mathieu Corteel, Propédeutique à l'évaluation des agencements entre humains et intelligences artificielles, LIEPP Policy Brief, n°81, 2025-09-30.

Introduction

Depuis l'émergence des intelligences artificielles LLM (Large Language Models)[1] dans le domaine public, les perspectives de croissance de tous les secteurs d'activité ont entraîné une spéculation économique et un engouement généralisé des politiques publiques pour un marché qui pourrait atteindre 1811,75 milliards de dollars d'ici 2030[2]. On parle d'un changement général des systèmes d'information et de communication impactant actuellement 400 millions d'utilisateurs, et qui pourrait toucher plus de 700 millions de personnes d'ici 2030 . Des experts indiquent que prochainement 90% du contenu d'internet sera généré par IA[3] alors même que la navigation sur internet sera entièrement assistée par IA[4]. Au même moment, le FMI annonce que 60% des emplois des pays développés seront en partie automatisés et transformés par l'IA (IMF, 2024) pour un gain de productivité global espéré de 7% en dix ans, selon une estimation de Goldman Sachs[5], qui pourrait aller jusqu'à 15%, si on en croit les prévisions de PwC[6].

Les prévisions de croissance ainsi que la démultiplication des usages laissent penser que ces technologies sont d'ores et déjà intégrées à l'ensemble de nos activités humaines. Cependant, le processus d'implémentation des IA au sein des organisations ne fait que commencer et se confronte à certaines limites. Pour interroger ces limites, ce Policy brief propose un cadre d'évaluation reposant sur les conditions de possibilité selon lesquelles les IA sont appelées à s'hybrider aux humains. Il reprend de manière synthétique les analyses développées dans l'ouvrage « Ni dieu ni IA » (Corteel, 2025). La démarche proposée ici consiste à interroger les IA au regard des agencements épistémiques, c'est-à-dire nos systèmes de croyances individuels et collectifs, dans lesquelles elles sont appelées à s'hybrider avec l'humain.

Difficultés d'implémentation au sein des organisations

1.1. Quelques erreurs à éviter

Force est de constater que l'IA LLM n'est pas nécessairement synonyme d'une intégration réussie au sein des organisations. Comme l'indique une récente étude de la RAND corporation (Ryseff, 2023), menée à partir d'entretiens auprès de 65 experts en science des données, le risque d'échec d'implémentation est à l'heure actuelle très élevé. Certains l'estiment autour de 70-80%. Mais, quelles sont les raisons de tels échecs ? L'étude de la RAND liste à ce propos cinq erreurs qu'il convient d'éviter : (1) la mauvaise compréhension de l'outil par les acteurs ; (2) le manque de données nécessaires à l'entraînement des IA au sein des organisations ; (3) le piège de la course à l'innovation, qui pousse à utiliser le dernier modèle d'IA plutôt que celui qui est stabilisé dans l'organisation ; (4) le manque d'infrastructures de données (data centers) au sein des organisations ; enfin (5) le fait de vouloir appliquer l'IA à des tâches trop difficiles à résoudre pour cette dernière.

L'étude de la RAND met ainsi en évidence que les échecs d'implémentation proviennent avant tout de problèmes organisationnels et épistémiques que l'on peut résoudre en favorisant l'alignement des compétences techniques avec les besoins spécifiques des métiers impliqués. L'IA doit en ce sens répondre à des besoins concrets d'usage et de connaissance inhérents à l'organisation.

1.2. Échelle d'agencements entre humains et IA

Une seconde étude, menée par l'équipe de Diyi Yang et Erik Brynjolfsson (2025) à Stanford, met en évidence les préférences des professionnels dans la répartition des tâches entre humains et IA au sein des organisations. En comparant les souhaits d'automatisation de certaines tâches, exprimées par 1500 professionnels, à la faisabilité technologique définie par 52 experts en IA, l'étude indique que la motivation principale des usagers porte sur une automatisation des tâches répétitives jugées stressantes, avec la volonté de se concentrer sur les tâches à forte valeur ajoutée (jugement, créativité et communication interpersonnelle). Le rapport introduit une échelle d'agencement entre humains et IA, la Human Agency Scale afin de mieux définir les attentes en termes de répartition des tâches :

- H1 : l'IA est totalement autonome dans la tâche à effectuer.

- H2 : l'IA est responsable de l'exécution de la tâche avec une supervision humaine minimale.

- H3 : l'humain et l'IA collaborent tout au long de la tâche à parts égales.

- H4 : l'humain est responsable de l'exécution de la tâche avec différents niveaux d'aide de l'IA.

[1] Les modèles d'IA LLM (Large Language Models) ou génératifs sont des réseaux de neurones profonds, généralement basés sur l'architecture transformer et comprenant des milliards de paramètres (parfois plus de 100 milliards). Ils sont capables de générer du texte, ou pour certains modèles spécialisés, des images, à partir d'une entrée appelée « prompt ». Le prompt est d'abord découpé en unités élémentaires appelées « tokens », qui sont converties en vecteurs numériques grâce à une couche d'embedding lexicale. Ces vecteurs sont ensuite traités à travers plusieurs couches du réseau transformer pour générer la sortie finale.»

[2] https://www.statista.com/forecasts/1449844/ai-tool-users-worldwide

[3] C'est ce que l'experte en IA Nina Schick a récemment affirmé dans plusieurs médias.

[4] https://www.nytimes.com/2025/07/11/technology/personaltech/ai-internet-browser-dia.html

[5] https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

[6] https://www.pwc.com/gx/en/news-room/press-releases/2025/ai-adoption-could-boost-global-gdp-by-an-additional-15-percentage.html

- H5 : l'implication humaine est essentielle dans l'exécution des tâches.

Selon ladite étude, 45 % des professions étudiées optent pour une hybridation H3, d'égal à égal entre humain et IA, alors que seules quelques professions (éditeur, mathématicien, ingénieur aérospatial etc.) ont choisi un agencement H5 où l'humain est entièrement responsable de l'exécution des tâches. Pour ainsi dire, la majorité des travailleurs interrogés ont exprimé le souhait d'engager une délégation étendue de leur activité, pour laisser plus de place au jugement, à la créativité et aux activités relationnelles. Toutefois, les perspectives de développement de l'IA empiètent aujourd'hui sur lesdites activités à forte valeur ajoutée, telle que la résolution de problèmes éthiques (Savulescu et Maslen, 2015), les processus d'innovation et de création (*Wei et al.*, 2022), ou encore la communication interpersonnelle (Matei, 2025). Comment donc préserver et valoriser la place du jugement humain dans l'agencement avec l'IA au sein des organisations ?

Anthropologie sceptique des agencements

Comme l'indiquent Henri Bergeron et Patrick Castel, dans leur dernier ouvrage « L'Organocène » (2025), nous vivons actuellement dans une société saturée d'organisations qui façonnent l'ensemble des transformations sociales et technologiques. L'IA n'y échappe pas. Son intégration au sein de nos activités organisationnelles ne se décrète ni par le seul progrès technologique ni par l'offre du marché ni par la volonté des politiques publiques, mais s'opère dans des compromis internes aux organisations, au croisement de leurs normes, de leurs pratiques, de leurs routines collectives et de leurs nouveaux arbitrages sur la répartition des tâches entre humains et IA.

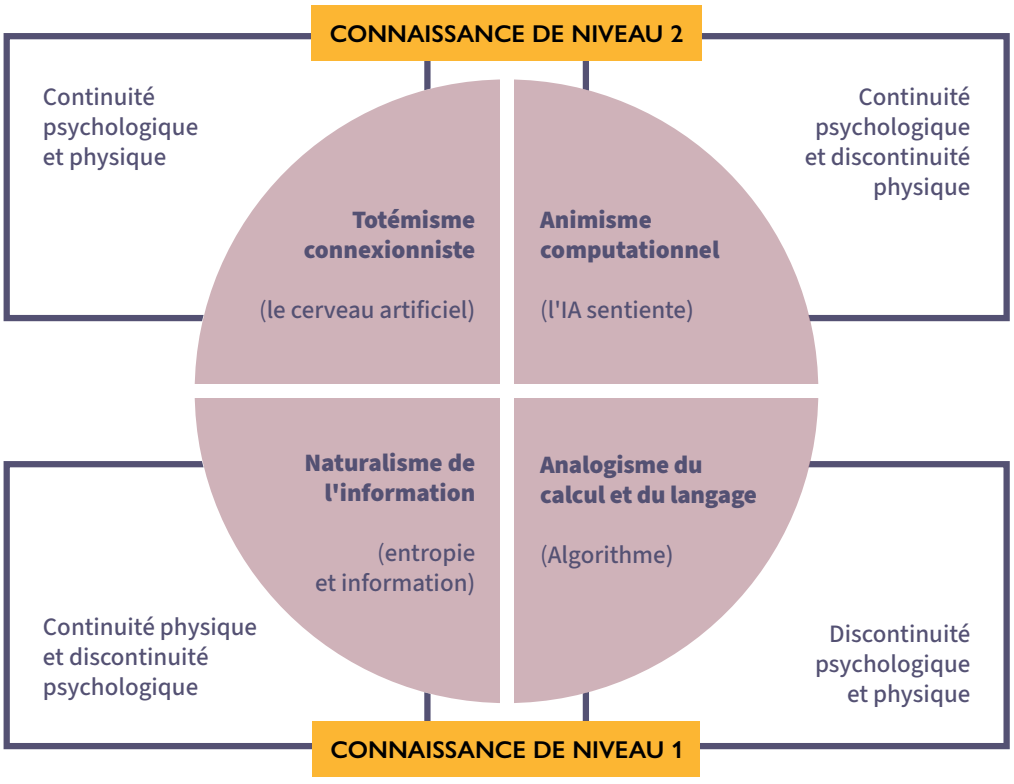
Toutefois, étant donné qu'il s'agit d'une technologie impactant directement la pensée de l'usager, une compréhension a priori de l'outil et des modalités de connaissance qui l'entourent doit servir de propédeutique à l'évaluation des IA au sein des organisations.

2.1. Modes d'existence des IA

Une telle connaissance a priori implique, selon Gilbert Simondon (1989), de décrire le mode d'existence propre de l'IA, c'est-à-dire de saisir son processus de concrétisation à travers l'ensemble des agencements réticulaires qu'elle noue avec l'humain. Appréhender ainsi la complexité de l'IA implique d'en circonscrire les liens anthropologiques externes que nous nouons avec ainsi que limites de calculabilité, de décision et d'action internes à la machine.

Dans « Ni dieu ni IA », les agencements entre humains et IA sont décrits à partir des rapports de continuité et de discontinuité physiques et mentales répartis selon deux ordres de connaissances dans lesquels s'hybrident humains et IA : (1) les connaissances de niveau 1 formelles (Cn1 = forme, calcul, syntaxe, combinatoire, ordre, rangement, logistique etc.) et (2) les connaissances de niveau 2 informelles (Cn2 = contextualisation, perception, attention, apprentissage, soin, création etc.). En s'inspirant des ontologies de Philippe Descola (2005), ces agencements ont pour vocation d'interroger les liens anthropologiques que nous nouons avec

Figure 1 : Tableau anthropologique des agencements entre humains et IA (Corteel, 2025)



les IA afin de discerner la part d'illusion qui peut apparaître chez l'usager qui projette du sens sur les effets de la combinatoire.

Incomplétude et limites épistémologiques de l'IA

L'argumentation développée dans « Ni dieu ni IA » propose ainsi de questionner ces agencements à partir des limites de l'IA en revenant sur un principe fondamental de la philosophie des mathématiques : tout est quantifiable, seulement, tout n'est pas calculable. Autrement dit, on peut attribuer une valeur numérique à toute chose et l'analyser au moyen d'un algorithme, mais ce n'est pas pour autant que l'on parviendra à comprendre tout ce qui est à l'œuvre dans le monde. Les vérités mathématiques sont plus vastes que ce qui est démontrable, le social est plus vaste que ce qui est modélisable.

3.1.1. Indécidabilité et incalculabilité

Dès les années 1930, le théorème d'incomplétude de Kurt Gödel (1931) mettait en évidence qu'il existe des énoncés mathématiques vrais mais impossibles à démontrer à partir des axiomes des mathématiques (Cassou, 2007). Autrement dit, il existe dans les mathématiques une part des énoncés qui reste indémontrable même si vraie du point de vue formel. On ne peut donc pas tout coder ni tout démontrer à partir d'un algorithme. Comme l'a théorisé Allan Turing (1937), il existe des programmes de démonstration qui ne s'arrêtent jamais, mais dont on ne peut pas prouver qu'ils s'arrêteront ou ne s'arrêteront jamais (Entscheidungsproblem). Ainsi, incomplétude et incalculabilité impliquent qu'une part de la pensée mathématique échappe à l'algorithmie. Ce qui libère une partie de la pensée humaine de l'IA.

3.1.2. Limites épistémologiques des données

De plus, il convient de considérer les limites de l'induction au principe du paramétrage de l'IA. Comme l'indique très justement Sabina Leonelli (2019), la construction des big data servant à l'entrainement des IA, se confrontent à quatre limites épistémologiques majeures. Ces dernières sont une source problématique de fausses corrélations (ou hallucinations) qui amoindrissent la fiabilité de nos IA :

(1) L'incommensurabilité des bases de données : comme les méthodes de classification des données se modifient avec le temps, mélanger des données issues d'une ancienne base de données avec des données récentes peut en effet générer de fausses corrélations.

(2) Les variations dans la collecte des données : les instruments et techniques utilisées pour collecter des données impactent la qualité des données récoltées et donc le résultat obtenu. Pour ainsi dire, toutes les données ne se valent pas.

(3) L'échantillonnage ou la sélection des données : on ne dispose que de données partielles. Un jeu de données, même s'il repose sur un grand nombre d'éléments et une méthode robuste, constitue toujours un échantillon de la réalité. La construction des données dépend de choix de représentation variables.

(4) La véracité des données : on peut corrompre des données pour améliorer leur valeur marchande, leur intérêt scientifique ou leur influence politique. Cela pose le problème de la neutralité dans la production des données, leur échange et leur circulation.

3.1.3. L'IA dégénérative ou les limites de l'apprentissage machine

Les modèles actuels d'IA LLM apprennent, ou du moins s'améliorent, par le fait d'une adaptation de leur paramétrage aux données qu'elles analysent. La particularité de ces IA LLM vient de la taille des paramètres. L'idée étant que si on étend le paramétrage de manière indéfinie on augmente les performances de manière indéfinie[7] (Brown *et al.*, 2020). La taille du modèle de ChatGPT.4.0 est actuellement d'environ 500 milliards de paramètres et elle ne fait que croître. Seulement, ni la démultiplication indéfinie du nombre de neurones paramétrés, ni l'accumulation étendue de toutes les données possibles ne permettent d'affirmer que l'on atteindra l'IA générale capable de dépasser la pensée humaine.

Les IA LLM se heurtent de fait à un mur de performance. Selon Peter Coveney et Sauro Succi (2025), la loi d'échelle appliquée à l'amélioration des LLM montre que leur progression s'avère limitée par un coût exponentiel en taille, données et énergie : chaque division par dix du taux d'erreur requiert une multiplication des ressources d'apprentissage (paramètres et données) par 10^{10}. Cette même division par dix du taux d'erreur implique une augmentation par 10^{20} de la consommation énergétique. Les IA LLM se confrontent donc à un mur de rendement manifeste qui reporte le progrès des modèles sur le *fine-tuning*[8] et le travail du clic pour nettoyer et affiner les paramètres et les

[7] Entre ChatGPT.2.0 et ChatGPT.3.0 on est passé d'un modèle doté de 1,5 milliards de paramètres à un modèle doté de 175 milliards de paramètres.

[8] Méthode d'apprentissage machine faite avec la supervision du programmateur pour appliquer l'IA à des tâches spécifiques ou corriger des erreurs identifiées dans certaines tâches.

données. Chaque amélioration de précision par *deep learning*[9] devient exponentiellement plus coûteuse. Ce qui implique à terme une stagnation des IA LLM et leur inadéquation à des tâches sensibles.

On assiste en même temps à une explosion des corrélations fallacieuses à mesure que les jeux de données s'élargissent, favorisant ainsi l'émergence de liens artificiels qui perturbent la connaissance (Calude et Longo, 2017). Ainsi, selon ces chercheurs, seule une réorientation vers la modélisation scientifique, basée sur des hypothèses explicites et la capacité à filtrer les corrélations pertinentes, permettra de sortir de cette trajectoire de « dégénérescence » où l'accroissement des données engendre plus d'erreurs que de progrès ; produit de ce qu'il convient d'appeler des « IA dégénératives ».

3.2. Favoriser les processus de co-construction

L'incomplétude et les limites de l'IA permettent de postuler qu'une part non-codifiable de notre pensée et de notre monde échappe inévitablement à l'IA (Corteel, 2025). En reconnaissant cette incomplétude, il s'agit d'éviter de déléguer à l'IA ce qui est incalculable et indécidable comme les décisions de nature éthique, la communication interpersonnelle ou bien la créativité. Dans l'ouvrage, ce raisonnement est appliqué aux agencements pragmatiques mêlant humain et IA selon deux catégories empruntées au neurobiologiste Francisco Varela : (1) les agencements allopoïétiques qui sont produits par autre chose qu'eux-mêmes et qui produisent autre chose qu'eux-mêmes ; et (2) les agencements autopoïétiques qui se génèrent eux-mêmes par l'ajustement de leur propre organisation interne en fonction de l'influence d'un milieu externe.

3.2.1. Éviter les agencements allopoïétiques au sein des organisations

L'objectif de cette démarche sceptique est de rappeler que l'IA ne fait que moduler, déplacer, hiérarchiser et ordonner la position de symboles qui n'ont aucune signification pour la machine elle-même. L'IA est allopoïétique. Les 0 et les 1 que l'IA manipule n'ont aucun sens pour l'IA. La sémantique nous appartient en propre. Aussi, il convient de valoriser dans le travail humain les tâches sémantiques pourvoyeuses de sens, tout en déléguant les tâches combinatoires à l'IA. Si on confond ces niveaux de connaissance, l'IA peut nous entraîner dans des paradoxes pragmatiques. Prenons par exemple un agencement allopoïétique humain/IA au sein d'une école : on a du côté de l'étudiant une utilisation naïve « prompt-clic » de l'IA pour générer son devoir, sa dissertation ou son commentaire, et du côté de l'enseignant une évaluation effectuée au moyen de l'IA pour définir les notations. On entre alors dans un agencement aliénant qui est produit par autre chose que lui-même et qui produit autre chose que lui-même.

La finalité de l'exercice atteinte et corrigée en un prompt et un clic, annihile tout le processus de co-construction, c'est-à-dire la production collective du sens commun à l'école. Pour contrer cela, peut-être ne s'agit-il pas tant de corriger l'IA que de travailler avec en favorisant la co-construction du sens commun avec des IA adaptées à l'institution. L'idée d'implémenter par exemple des IA socratiques (Lara et Deckers, 2020) amenant l'étudiant à se questionner plutôt qu'à obtenir une réponse immédiate, semble en ce sens une perspective à envisager.

3.3. Développer de nouvelles méthodes d'évaluation

Afin donc de concevoir, d'expérimenter et d'évaluer les agencements entre humains et IA dans des contextes particuliers, il serait intéressant de développer une analyse de ce que John Rawls nomme « l'équilibre réfléchi » (Rawls, 2009), c'est-à-dire le processus conduisant à un état de cohérence du jugement obtenu par ajustement entre des principes et des croyances particulières. Il s'agirait d'étudier le processus de réflexion de l'agent en dialogue avec l'IA qui conduit à la stabilisation de son jugement personnel à partir de deux types d'équilibre réfléchi : l'équilibre réfléchi de type 1 (ER1) qui conduit au renforcement des convictions de l'agent à partir du dialogue avec l'IA ; et l'équilibre réfléchi de type 2 (ER2) qui conduit l'agent à modifier son jugement à partir du dialogue avec l'IA.

En parallèle, il faudrait circonscrire ce qui perturbe l'équilibre réfléchi et le transforme en déséquilibre irréfléchi (DI). Comme le montrent les études de David Chavalarias (2020) sur les réseaux sociaux, deux types de biais cognitifs majeurs conduisent à la manipulation de l'opinion : (1) le biais de confirmation qui transforme l'équilibre réfléchi de type 1 en un déséquilibre irréfléchi de type 1 (DI1) par la formation de bulles de filtres et de chambres d'écho ; et (2) le biais de négativité, qui favorise un déséquilibre irréfléchi de type 2 (DI2) en jouant sur les passions tristes (peur, angoisse etc) grâce au relai et à la diffusion réactive d'informations négatives dans le réseau.

En 2014, le scandale de Facebook-Cambridge Analytica a montré la puissance de ce type d'agencement sur le choix social. De fait, la plateforme logicielle Ripon est parvenu à orienter le choix en faveur du Brexit, de Trump et de Bolsonaro en modélisant la tournure d'esprit des votants à partir des données volées à 87 millions de comptes utilisateurs

[9] Méthode d'apprentissage machine qui utilise des réseaux de neurones artificiels profonds pour apprendre automatiquement, c'est-à-dire générer des modèles ou patterns à partir du paramétrage des neurones sur de grandes quantités de données.

Facebook. Maintenant que les réseaux sociaux intègrent des IA orientées idéologiquement dans nos espaces numériques, il est urgent de participer à la recherche sur les modalités du jugement afin de se prémunir des ingérences pouvant porter atteinte au choix social et nous entrainer dans des agencements allopoïétiques aliénants.

Une telle analyse pourrait se faire sur un corpus de dialogues collecté après un appel à participation (dont l'échelle pourrait varier selon les objectifs : organisation, échelle locale ou nationale) à partir d'une démarche déontologique et d'une méthode d'évaluation reposant sur l'essai randomisé (Cartwright et Hardie, 2012). La perspective expérimentale serait de faire dialoguer des agents humains avec des chatbots à propos d'un des devoirs prima facie ou devoir moral intuitif (Ross, 1960), afin de voir si l'IA va renforcer la conviction de la personne à ce sujet ou bien l'amener à changer de position (ER1/DI1 ou ER2/DI2). Il s'agit ici d'analyser les variations de l'attitude propositionnelle à l'égard des devoirs prima facie. On pourrait ainsi établir une évaluation des IA génératives standards et des IA génératives socratiques à partir dudit corpus.

Conclusion

Actuellement, les IA LLM génèrent une crise interne à nos organisations qui touche à la dimension épistémique de l'ensemble de nos activités individuelles et collectives. La posture sceptique permet à cet égard de douter de manière heuristique des performances de l'IA promues par les concepteurs et d'adopter une approche pragmatique dans les usages.

C'est pourquoi cette propédeutique des agencements propose d'orienter les politiques publiques vers une évaluation de l'implémentation de l'IA dans les organisations en considérant (1) les limites épistémologiques des IA ; (2) la complexité des agencements et des liens de continuité/discontinuité physiques et mentaux qui se nouent à travers deux niveaux de connaissance partagés entre l'IA et l'usager ; enfin (3) évaluer les processus de dotation de sens dans les activités humaines en distinguant l'allopoïétique de l'autopoïétique, c'est-à-dire en distinguant la combinatoire dénuée de sens du processus de dotation de sens inhérent aux activités humaines qu'il convient de valoriser.

Figure 2 : Tableau Équilibre réfléchi/Déséquilibre irréfléchi dans le dialogue humain-IA

Catégorie	Type	Description du processus	Mécanisme clé	Résultat pour l'agent
Équilibre réfléchi	ER1 (Renforcement)	Dialogue avec IA consolidant les jugements de manière cohérente	Dialogue réflexif	Le jugement de l'agent est confirmé après avoir été mis à l'épreuve
	ER2 (Modification)	Le dialogue avec IA exposant l'agent à de nouvelles informations, des arguments pertinents ou des incohérences dans son propre raisonnement, le conduisant à ajuster ou changer son jugement	Dialogue ouvert à la nouveauté et à la correction	Modification éclairée du jugement. L'agent adopte une nouvelle position de manière réfléchie et cohérente
Déséquilibre irréfléchi	DI1 (Enfermement)	L'interaction avec l'IA crée une «bulle de filtre» ou une «chambre d'écho» qui ne présente à l'agent que des informations confirmant ses croyances initiales	Biais de confirmation. L'IA exploite la tendance humaine à ne chercher que ce qui valide ses opinions	Enfermement dogmatique. La conviction est renforcée sans réflexion ni confrontation, menant à une radicalisation
	DI2 (Manipulation)	L'IA modifie le jugement de l'agent en le surexposant à des informations négatives et anxiogènes, déclenchant une réaction émotionnelle plutôt qu'une analyse rationnelle	Biais de négativité. L'IA joue sur les « passions tristes » (peur, angoisse) pour orienter le jugement	Manipulation émotionnelle. L'agent change d'avis non par conviction, mais par peur ou anxiété

Références

A. Casilli, *En attendant les robots, Paris*, Seuil, 2019.

A. Matei, « *'Hey man, I'm so sorry for your loss': should you use AI to text?* », The Guardian, 30/07/2025.

A. Turing, "On computable numbers, with an application to the entscheidungsproblem". *Proceedings of the London Mathematical Society*, s2-42(1), 1937, pp. 230-26.

Brown *et al.*, "*Language Models are Few-Shot Learners*", arXiv:2005.14165, 2020.

C.S. Calude & G. Longo, "*The deluge of spurious correlations in big data*", Foundations of Science, 22(3), 2017, pp. 595–612.

Coveney, P.V. & Succi, S. (2025) "The wall confronting large language models", arXiv preprint arXiv:2507.19703

D. Chavalarias, *Toxic Data, comment les réseaux manipulent nos opinions*, Paris, Flammarion, 2022.

F. Lara, J. Deckers, "*Artificial Intelligence as a Socratic Assistant for Moral Enhancement*", Neuroethics, 2020(13), pp. 275-287.

Grand view research, Artificial Intelligence Market Size, Share & Trends Analysis Report, 2025 – 2030.

G. Simondon, *Du mode d'existence des objets techniques*, Paris, Aubier, 1989.

H. Bergeron et P. Castel, *L'Organocène*, Paris, Presses de Sciences Po, 2025.

IMF, *Gen-AI: Artificial Intelligence and the Future of Work*, Janvier 2024.

J. Rawls, *Théorie de la justice*, trad. C. Audard, Paris, Points, 2009, p. 74.

J. Ryseff *et al.*, *The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed, Avoiding the Anti-Patterns of AI*, Rand Corporation, 2023.

J. Savulescu, H. Maslen, *"Moral Enhancement and Artificial Intelligence: Moral AI?"*, in J. Romportl, E. Zackova, J. Kelemen (eds) *Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics*, vol 9, Springer, 2015.

K. Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme", I, *Monatshefte für Mathematik und Physik*, 38(1), 1931, pp. 173-198.

N. Cartwright & J. Hardie, *Evidence-based policy : a practical guide to doing it better.* Oxford University Press, 2012.

M. Corteel, *Ni dieu ni IA, une philosophie sceptique de l'intelligence artificielle*, Paris, La découverte, 2025.

P. Cassou-Noguès, *Les démons de Gödel : logique et folie*, Paris, Le seuil, Science ouverte, 2007.

P. Descola, *Par-delà nature et culture*, Paris, Gallimard, Bibliothèque des Sciences Humaines, 2005.

S. Leonelli, *La recherche scientifique à l'ère des big data, cinq façons dont les big data nuisent à la science et comment la sauver*, Paris, Mimesis, 2019.

W. D. Ross, *Foundations of ethics: the Gifford lectures delivered in the University of Aberdeen*, 1935-36. Clarendon Press, 1960.

Wei *et al.*, "*Emergent Abilities of Large Language Models*", arXiv:2206.07682, 2022.

D. Yang, E. Brynjolfsson *et al.*, *"Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce"*, arXiv:2506.06576.

Le LIEPP (laboratoire interdisciplinaire d'évaluation des politiques publiques) est un laboratoire d'excellence (Labex) distingué par le jury scientifique international désigné par l'Agence nationale de la recherche (ANR).

Il est financé dans le cadre des investissements d'avenir de l'IdEx Université Paris Cité *(ANR-18-IDEX-0001).*

Pour recevoir les prochains échos du LIEPP et rester informé de nos activités, merci d'envoyer un courriel à : liepp@sciencespo.fr

Directrice de publication :
Anne Revillard
Edition et maquette :
Andreana Khristova
Evane Grossemy

Sciences Po - LIEPP
27 rue Saint Guillaume
75007 Paris - France
+33(0)1.45.49.83.61