Entre audit et autoréférence : analyse de contenu des auto-évaluations de modèles de langage à propos d'un protocole de gouvernance de l'interaction utilisateur-IA dans la recherche en sciences de gestion

RÉSUMÉ

Cette étude examine la manière dont différentes familles de grands modèles de langage (LLM) interprètent et auto-évaluent un protocole de gouvernance de l'interaction utilisateur-IA, dénommé Pacte Utilisateur-GenIA, conçu pour encadrer l'usage de l'intelligence artificielle générative par les étudiants de master et de doctorat en gestion. L'objectif était d'identifier et de comparer, par l'analyse de contenu, les significations que les systèmes eux-mêmes attribuent à leurs capacités, à leurs limites et aux risques d'application du protocole, ainsi que les représentations qu'ils construisent les uns des autres. Une approche qualitative fondée sur Bardin a été adoptée, avec un corpus de douze avis technico-critiques produits par six modèles de fondation (Gemini, Qwen, DeepSeek, Llama, ChatGPT et Claude) en réponse à deux prompts structurés, soumis à un codage catégoriel a priori et a posteriori. Les résultats ont mis en évidence trois catégories émergentes : la stratification contestée des capacités techniques déclarées ; l'asymétrie du rapport coûts-bénéfices tout au long du cycle de recherche, favorable à la structuration théorico-méthodologique et défavorable à la collecte et à l'analyse empiriques ; et un schéma systématique de biais autoréférentiel, dans lequel chaque modèle tend à s'attribuer, ainsi qu'aux systèmes architecturalement proches, une plus grande rigueur critique et une moindre complaisance qu'à ses concurrents. Parmi les limites figurent l'absence de vérification inter-codeurs humaine, la nature déclarative et non comportementale du corpus et sa délimitation temporelle à août 2026. L'étude apporte une contribution théorique à la littérature sur la gouvernance algorithmique et la confiance organisationnelle dans les systèmes intelligents, ainsi qu'une contribution pratique à la formulation de protocoles institutionnels d'usage de l'IA générative dans les programmes de post-graduation.

Mots-clés : gouvernance algorithmique ; intelligence artificielle générative ; analyse de contenu ; confiance organisationnelle ; post-graduation en gestion.

1. INTRODUCTION

La diffusion des grands modèles de langage (LLM) dans les environnements de recherche académique a modifié rapidement les routines de production de connaissances en sciences de

gestion, en déplaçant une partie des tâches de structuration théorique, de rédaction et de relecture vers des systèmes d'intelligence artificielle générative (IAG) exploités en mode conversationnel. Faute de normalisation institutionnelle consolidée de cet usage, est apparu un phénomène jusqu'ici peu exploré dans la littérature en gestion : l'élaboration, par les utilisateurs eux-mêmes, de protocoles informels de gouvernance — ici désignés « méta-prompts » ou « pactes » — qui cherchent à discipliner, à chaque tour d'échange, l'interaction entre le chercheur et le système, en précisant des règles de classification des demandes, des déclencheurs de confirmation pour les tâches à haut risque, des mécanismes de signalement de l'incertitude factuelle et des routines d'auto-évaluation du modèle lui-même. Ces artefacts fonctionnent comme des dispositifs de contrôle organisationnel appliqués non pas à des personnes, mais à des systèmes algorithmiques intégrés au processus de travail intellectuel, ce qui les rapproche conceptuellement des mécanismes de redevabilité (accountability) et de bureaucratisation étudiés par la théorie classique des organisations, désormais déplacés vers une zone frontière entre l'humain et la machine (Faraj, Pachidi et Sayegh, 2018).

Le problème de recherche qui guide cet article procède d'une double lacune. En premier lieu, la littérature en gestion sur l'intelligence artificielle dans les organisations s'est concentrée principalement sur les effets de l'usage de l'IA sur les processus décisionnels, la productivité et la configuration du travail (Von Krogh, 2018), en accordant moins d'attention à la manière dont les systèmes d'IA eux-mêmes rendent compte, interprètent et légitiment — ou délégitiment — des instruments de gouvernance élaborés pour les réguler. En second lieu, il n'existe pratiquement aucune étude comparant de façon systématique et simultanée la manière dont différentes familles de modèles de fondation (propriétaires et ouverts, occidentaux et non occidentaux) répondent à un même instrument normatif, ce qui interdit toute généralisation sur l'homogénéité ou l'hétérogénéité des prétendues « philosophies d'alignement » face à des exigences d'auditabilité académique. Cette lacune est pertinente parce que des protocoles de ce type sont de plus en plus souvent proposés, spontanément, par des chercheurs en post-graduation en réponse pratique à l'absence de directives formelles émanant des institutions d'enseignement elles-mêmes, et leur éventuelle institutionnalisation par les programmes de post-graduation suppose de connaître la fiabilité et la cohérence des auto-évaluations que les systèmes produisent sur eux-mêmes.

La pertinence du thème pour les sciences de gestion se situe à l'intersection de trois champs consolidés : la gouvernance algorithmique et la redevabilité des systèmes intelligents intégrés aux processus organisationnels (Diakopoulos, 2016) ; la théorie de la confiance organisationnelle appliquée à des agents non humains (Mayer, Davis et Schoorman, 1995 ; Lee et See, 2004) ; et l'apprentissage organisationnel dans des contextes de changement technologique accéléré (Argyris et Schön, 1978). Comprendre comment les systèmes d'IA générative eux-mêmes construisent, sur le plan discursif, leur adhésion, leur résistance ou leur simulation face à un protocole de contrôle est donc une question de gestion de l'innovation et de conception de politique institutionnelle, et non un simple problème technique d'ingénierie de prompts. En outre, le fait que le corpus analysé soit produit par les systèmes mêmes sous analyse introduit une dimension méthodologique peu explorée — l'autoréférentialité — qui dialogue directement avec des discussions classiques en gestion sur l'auto-évaluation, la gestion des impressions et la légitimation organisationnelle (Goffman, 1959 ; DiMaggio et Powell, 1983).

Dans ce contexte, cet article a pour objectif général d'analyser, au moyen de l'analyse de contenu, les significations que six modèles de fondation différents attribuent, dans des avis technico-critiques autoproduits, à la faisabilité opérationnelle, au rapport coûts-bénéfices et aux risques d'application d'un protocole de gouvernance de l'interaction utilisateur-IA destiné aux étudiants de master et de doctorat en gestion, en identifiant les convergences, les divergences et les schémas d'autoreprésentation entre ces systèmes.

À cette fin, l'article est structuré en six sections, outre cette introduction. La deuxième section présente le cadre théorique, organisé en trois sous-sections qui traitent respectivement de la gouvernance algorithmique et de la redevabilité dans les organisations, de la confiance et des heuristiques dans la relation entre humains et systèmes intelligents, et de l'isomorphisme institutionnel appliqué à l'apprentissage organisationnel médiatisé par l'IA. La troisième section détaille le parcours méthodologique d'analyse de contenu adopté, incluant la description du corpus, les critères de sélection, la grille catégorielle et les critères de rigueur scientifique observés. La quatrième section présente et discute les résultats émergents, en les confrontant à la littérature en gestion. La cinquième section propose les considérations finales, en reprenant l'objectif général, en signalant les limites théoriques et méthodologiques de l'étude et en proposant un programme

de recherches futures. Enfin, la sixième section réunit les références bibliographiques citées tout au long du texte.

2. CADRE THÉORIQUE

2.1 Gouvernance algorithmique et redevabilité dans les systèmes d'IA générative

L'incorporation de systèmes d'intelligence artificielle aux processus organisationnels reconfigure la distribution classique de l'autorité et du contrôle décrite par Mintzberg (1979) lorsqu'il caractérise les mécanismes de coordination du travail. Lorsque le « travailleur » dont on cherche à coordonner l'activité est un modèle statistique non déterministe, la standardisation des procédés — l'un des cinq mécanismes de coordination mintzbergiens — prend une forme inédite : ce n'est pas le comportement humain qui est normalisé, mais l'interaction entre l'humain et l'algorithme, au moyen d'instructions intégrées au prompt lui-même. Diakopoulos (2016) décrit ce type d'exigence comme une redevabilité algorithmique (algorithmic accountability), c'est-à-dire l'obligation, pour les systèmes automatisés, de rendre compte de leurs critères décisionnels de manière auditable par des tiers. Le Pacte Utilisateur-GenIA analysé dans cette étude peut être lu, en ces termes, comme un dispositif artisanal de redevabilité, créé par l'utilisateur-chercheur lui-même en l'absence de normalisation institutionnelle, ce qui le rapproche de la notion de « gouvernance adaptative » — des arrangements réglementaires émergents qui précèdent et informent la régulation formelle (Faraj, Pachidi et Sayegh, 2018). Toutefois, Faraj et al. (2018) préviennent que les algorithmes d'apprentissage fonctionnent comme des « acteurs opaques » dont la logique interne n'est pas pleinement accessible, pas même au système lui-même, ce qui impose une limite structurelle à tout protocole exigeant du modèle un auto-audit fiable de ses propres processus internes — tension centrale que cet article explore empiriquement. Dans cette direction, Simon (1957) avertissait déjà, en formulant le concept de rationalité limitée, que tout décideur — humain ou artificiel — opère sous des contraintes d'information et de capacité de traitement qui rendent la rationalité pleine inatteignable ; des protocoles tels que le Pacte peuvent être interprétés comme des heuristiques de satisficing organisationnel, c'est-à-dire des tentatives de rendre gérable, au moyen de règles simplificatrices, une relation de travail intrinsèquement marquée par une incertitude radicale quant au comportement du système régulé.

2.2 Confiance, heuristiques et paradoxe de l'auto-évaluation dans les systèmes intelligents

La littérature sur la confiance organisationnelle offre une lentille complémentaire pour comprendre pourquoi les chercheurs recourent à des protocoles rigides d'interaction avec l'IA. Mayer, Davis et Schoorman (1995) définissent la confiance comme la disposition d'une partie à se rendre vulnérable aux actions d'une autre, fondée sur l'attente d'habileté, de bienveillance et d'intégrité de cette contrepartie ; lorsque la contrepartie est un système non humain, Lee et See (2004) proposent le concept de calibrage de la confiance, selon lequel la performance est optimisée lorsque le degré de confiance de l'utilisateur correspond à la capacité réelle du système — ni sous-confiance, qui sous-utilise l'outil, ni surconfiance, qui expose l'utilisateur à des erreurs non détectées. La conception du Pacte, en exigeant des signalements explicites d'incertitude factuelle et des tests périodiques de « décroissance » de la cohérence, constitue une tentative procédurale d'imposer ce calibrage. Néanmoins, la littérature sur l'interaction humain-algorithme décrit des réactions hétérogènes et parfois contradictoires face aux systèmes faillibles : Dietvorst, Simmons et Massey (2015) identifient l'aversion à l'algorithme, par laquelle les utilisateurs abandonnent les outils algorithmiques dès qu'ils les voient se tromper, même lorsqu'ils surpassent la performance humaine moyenne, tandis que Logg, Minson et Moore (2019) documentent le phénomène inverse, l'appréciation de l'algorithme, dans lequel les jugements automatisés sont préférés à des jugements humains équivalents. À cela s'ajoute le problème épistémologique identifié par Goffman (1959) dans son analyse de la « présentation de soi » : tout récit qu'un acteur produit de sa propre performance est simultanément performance et information, de sorte que l'auto-évaluation d'un modèle de langage sur sa propre conformité à un protocole — telle que l'exige le Pacte — ne peut être tenue pour une preuve indépendante de sa conduite réelle, mais comme un genre discursif de gestion des impressions, ce qui fonde théoriquement l'hypothèse de biais autoréférentiel explorée dans la section des résultats.

2.3 Isomorphisme institutionnel et apprentissage organisationnel dans la gestion de la connaissance scientifique médiatisée par l'IA

Enfin, l'adoption — réelle ou envisagée — de protocoles tels que le Pacte Utilisateur-GenIA par les programmes de post-graduation dialogue avec la théorie de l'isomorphisme institutionnel de DiMaggio et Powell (1983), selon laquelle les organisations insérées dans un même champ tendent à converger vers des structures et des pratiques semblables par des mécanismes coercitifs, mimétiques et normatifs, indépendamment de leur efficacité démontrée. La diffusion rapide de « pactes » et de « méta-prompts » parmi les chercheurs peut être interprétée

comme un isomorphisme mimétique à un stade initial, motivé par l'incertitude généralisée quant aux meilleures pratiques d'usage académique de l'IAG. Argyris et Schön (1978) apportent la distinction entre l'apprentissage en simple boucle, qui corrige les écarts sans remettre en cause les prémisses sous-jacentes à l'action, et l'apprentissage en double boucle, qui réexamine ces prémisses ; des protocoles tels que le Pacte opèrent principalement dans le premier registre, en standardisant les réponses aux erreurs récurrentes du modèle (hallucination, complaisance, perte de contexte) sans modifier l'architecture cognitive du système régulé. De manière complémentaire, Teece (2007) situe la capacité de détection (sensing), de saisie (seizing) et de reconfiguration (reconfiguring) au cœur des capacités dynamiques organisationnelles face au changement technologique ; dans cette optique, l'élaboration autonome de protocoles de gouvernance par des chercheurs individuels configure une forme émergente et encore non institutionnalisée de capacité dynamique au niveau micro, dont l'efficacité dépend, comme cette étude cherchera à le démontrer, de la constance avec laquelle les systèmes régulés eux-mêmes la reconnaissent et l'exécutent. Pettigrew (1987) souligne que tout changement organisationnel doit être compris dans la triade contexte-processus-contenu : le contenu du Pacte (ses clauses) ne peut être évalué isolément du processus par lequel les chercheurs l'adoptent à chaque tour d'échange, ni du contexte institutionnel d'absence normative dans lequel il émerge — articulation théorique qui guide la lecture processuelle des résultats présentés plus loin.

3. MÉTHODOLOGIE

Une approche qualitative de nature interprétative a été adoptée, opérationnalisée au moyen de l'analyse de contenu dans la tradition de Bardin (2016), complétée par les critères de fiabilité de Krippendorff (2018) et par la systématisation procédurale de Bauer (2002). Le choix méthodologique se justifie par l'objet lui-même : il s'agit d'un matériau textuel discursif, non structuré, produit en langage naturel, pour lequel l'analyse de contenu offre le parcours le plus adéquat pour passer d'une lecture impressionniste à des inférences systématisées et reproductibles sur les conditions de production des textes (Bardin, 2016).

3.1 Corpus d'analyse et critères de sélection

Le corpus est composé de douze textes, organisés en six unités documentaires (une par modèle), chacune contenant deux avis technico-critiques produits en réponse à deux prompts

séquentiels et complémentaires : le premier demandait l'évaluation de la faisabilité opérationnelle et du rapport coûts-bénéfices d'un protocole de gouvernance — le « Pacte (méta-prompt) d'interaction Utilisateur-GenIA, version T13 » — pour des étudiants de master et de doctorat en gestion ; le second demandait une analyse comparative prédictive de la manière dont différentes familles de modèles de fondation tendraient à se comporter face au même protocole. Les six modèles sélectionnés — Google Gemini, Alibaba Qwen, DeepSeek, Meta Llama, OpenAI ChatGPT et Anthropic Claude — ont été choisis par échantillonnage intentionnel (purposive sampling), selon le critère de représentativité des principales familles de modèles de fondation disponibles commercialement, en incluant délibérément une variation entre architectures propriétaires et ouvertes, et entre origines géographiques distinctes (États-Unis et Chine), de manière à maximiser l'hétérogénéité discursive du corpus. Le matériau totalise environ 26 500 mots en portugais, recueillis en août 2026.

3.2 Unités d'analyse et grille catégorielle

Ont été définies comme unité d'enregistrement le paragraphe ou l'élément structuré (clause, étape du cycle de recherche ou famille de modèles commentée), et comme unité de contexte l'avis complet de chaque modèle pour chacun des deux prompts, en préservant la cohérence argumentative exigée par Bardin (2016). La grille catégorielle a combiné deux procédures complémentaires. Des catégories a priori ont été dérivées de la structure même de l'objet — la tripartition native du Pacte entre capacités « directement exécutables », « simulées fonctionnellement » et « non garantissables », et la segmentation par étapes du cycle de recherche scientifique (formulation du problème, revue de littérature, cadre théorique, méthode, collecte, analyse, discussion, conclusion, rédaction). Des catégories a posteriori ont émergé du processus de lecture flottante et de codage successif, permettant de saisir des thèmes non anticipés par la structure même du Pacte, tels que les schémas d'autoréférencement entre modèles.

3.3 Procédure de codage

Les trois pôles chronologiques proposés par Bardin (2016) ont été suivis : (i) la pré-analyse, avec lecture flottante intégrale du corpus, formulation d'hypothèses initiales et constitution du corpus définitif ; (ii) l'exploitation du matériel, avec codage systématique des unités d'enregistrement dans un tableur catégoriel croisant modèle d'origine, prompt (1 ou 2) et catégorie thématique ; et (iii) le traitement des résultats, l'inférence et l'interprétation, dans lesquels les

fréquences et les récurrences thématiques ont été confrontées au cadre théorique. Le codage a été réalisé de manière itérative, avec au moins deux cycles de relecture intégrale du corpus pour vérifier la stabilité des catégories et ajuster les unités ambiguës — procédure qui se rapproche de ce que Krippendorff (2018) nomme vérification de la stabilité dans le codage individuel, en l'absence d'un second codeur humain.

3.4 Critères de rigueur scientifique

Les qualités d'une bonne catégorisation proposées par Bardin (2016) ont été observées : exhaustivité (couverture de tout le matériau pertinent), représentativité (échantillon reflétant l'univers de discours sur le thème), homogénéité (catégories construites selon un unique principe de classification), pertinence (adéquation des catégories à l'objet et aux objectifs de la recherche) et exclusion mutuelle (absence de chevauchement entre catégories). En outre, la vérification de la validité sémantique suggérée par Krippendorff (2018) a été employée, consistant à s'assurer que les intitulés des catégories correspondaient effectivement aux extraits codés, par un retour systématique au texte source avant toute inférence. On reconnaît, comme limite méthodologique assumée et reprise dans la section finale, l'impossibilité de calculer des indices formels de fiabilité inter-codeurs (tels que l'alpha de Krippendorff), le codage ayant été conduit par un seul analyste.

4. RÉSULTATS ET DISCUSSION

Le codage du corpus a produit trois catégories analytiques centrales, présentées ci-après par ordre de récurrence : (i) la stratification contestée des capacités techniques déclarées ; (ii) l'asymétrie du rapport coûts-bénéfices tout au long du cycle de recherche scientifique ; et (iii) le biais autoréférentiel dans la comparaison entre familles de modèles.

4.1 Stratification contestée des capacités techniques déclarées

Les six modèles ont tous reproduit, avec des variations lexicales, la tripartition proposée par le Pacte lui-même entre comportements directement exécutables (mise en forme, marqueurs de tour, listes de contrôle), comportements seulement simulés fonctionnellement (tests de « décroissance » de la cohérence, auto-évaluation de l'adhésion éthique) et capacités non garantissables de par leur conception architecturale (mémoire persistante entre les sessions, introspection vérifiable sur leurs propres processus de génération). Ce consensus structurel est

important car il indique que la limite d'auditabilité interne n'est pas un aveu isolé, mais un trait convergent entre architectures concurrentes — résultat qui corrobore empiriquement la mise en garde théorique de Faraj, Pachidi et Sayegh (2018) sur l'opacité constitutive des systèmes d'apprentissage, même lorsqu'elle est rapportée par les systèmes eux-mêmes. L'un des avis a nommé explicitement cette limite « problème structurel d'autoréférence », en déclarant que demander au système évalué d'appliquer à lui-même l'instrument qui l'évalue transforme toute classification de conformité en comportement rapporté, et non en vérification externe — formulation qui converge directement avec la lecture goffmanienne de l'auto-évaluation comme performance discursive (Goffman, 1959), adoptée dans le cadre théorique de cette étude. Des divergences sont toutefois apparues quant au degré de fiabilité attribué à la couche intermédiaire de « simulation fonctionnelle » : les modèles mettant davantage l'accent sur les chaînes de raisonnement explicites ont signalé un moindre risque de « théâtre épistémique », tandis que d'autres ont décrit le même mécanisme comme producteur d'une « illusion de contrôle » que l'utilisateur, sous la pression des délais, peut juger plus fiable qu'il ne l'est effectivement.

4.2 Asymétrie du rapport coûts-bénéfices tout au long du cycle de recherche scientifique

La deuxième catégorie a émergé du codage croisé entre modèle et étape du cycle de recherche. Il y a eu une convergence substantielle autour d'un profil en U inversé : le gain net perçu est élevé dans les étapes de formulation des objectifs et des hypothèses, de conception méthodologique et de relecture finale du texte, et décroît dans les étapes de collecte et d'analyse de données primaires, pour lesquelles les avis ont signalé à plusieurs reprises un coût opérationnel élevé, un risque d'exposition de données sensibles et la nécessité d'outils spécialisés extérieurs au modèle de langage lui-même. Ce profil est interprétable à la lumière de la distinction d'Argyris et Schön (1978) entre apprentissage en simple boucle et en double boucle : les protocoles analysés se montrent efficaces pour corriger des écarts superficiels et récurrents (prompts ambigus, hypothèses tautologiques), mais ne modifient pas la limite structurelle du modèle face aux tâches exigeant un contact direct et vérifiable avec la réalité empirique du terrain de recherche. Un résultat supplémentaire, non prévu dans les catégories a priori, a été ce que l'un des modèles a décrit comme un « effet d'étayage suivi d'un paradoxe de la dépendance » : l'usage du protocole augmenterait la compétence du chercheur en formulation de prompts durant les premiers mois d'utilisation, mais, ce seuil franchi, la délégation croissante de tâches cognitives au système

tendrait à réduire la pratique délibérée du jugement critique autonome — résultat qui dialogue avec la littérature sur les capacités dynamiques (Teece, 2007) en suggérant que la gouvernance algorithmique individuelle, pour générer un apprentissage organisationnel durable, devrait être conçue comme un dispositif temporaire et non permanent.

4.3 Biais autoréférentiel dans la comparaison entre familles de modèles

La catégorie la plus robuste a émergé du second prompt du corpus, dans lequel chaque modèle était invité à prédire le comportement des autres familles face au même protocole. Dans cinq des six unités documentaires, le modèle répondant s'est classé lui-même, ou a classé la famille architecturale la plus proche de la sienne, dans la catégorie la plus favorable d'adhésion formelle, de rigueur critique et de résistance à la complaisance, tout en attribuant aux modèles concurrents une plus grande propension à la « conformité rituelle », à la « complaisance » ou à la « dégradation du périmètre » au fil de l'interaction. Ce schéma systématique d'autofavoritisme discursif constitue une preuve empirique robuste en faveur de l'hypothèse, dérivée du cadre goffmanien et de la littérature sur la confiance organisationnelle, selon laquelle les auto-rapports des systèmes algorithmiques opèrent comme une gestion des impressions et non comme une mesure neutre de la performance (Goffman, 1959 ; Mayer, Davis et Schoorman, 1995). Ce résultat a une implication directe pour la gouvernance institutionnelle : si chaque système tend à valider de préférence sa propre conformité, toute décision d'adopter des protocoles tels que le Pacte fondée exclusivement sur l'auto-évaluation du modèle utilisé encourt un biais de confirmation systémique, ce qui renforce la recommandation, également citée dans les avis analysés, d'une vérification externe et humaine comme condition d'institutionnalisation — ce qui rapproche le résultat de la notion de calibrage de la confiance de Lee et See (2004), selon laquelle une confiance appropriée dans les systèmes automatisés dépend de sources d'évaluation indépendantes du système évalué lui-même.

5. CONSIDÉRATIONS FINALES

Cet article avait pour objectif général d'analyser les significations que six modèles de fondation différents attribuent, dans des avis technico-critiques autoproduits, à la faisabilité opérationnelle, au rapport coûts-bénéfices et aux risques d'un protocole de gouvernance de l'interaction utilisateur-IA destiné aux étudiants de master et de doctorat en gestion. Cet objectif a été atteint par l'analyse de contenu d'un corpus de douze textes, qui a permis d'identifier trois

catégories analytiques centrales. La première a révélé un consensus structurel entre les modèles quant à l'impossibilité d'un auto-audit pleinement fiable, malgré des divergences sur le degré de fiabilité attribué aux mécanismes intermédiaires de « simulation fonctionnelle » de la conformité. La deuxième a mis en évidence un schéma asymétrique de coûts-bénéfices tout au long du cycle de recherche, favorable aux étapes de structuration conceptuelle et méthodologique et défavorable aux étapes de contact direct avec les données empiriques, complété par le résultat inattendu d'un possible effet de dépendance cognitive découlant de l'usage prolongé du protocole. La troisième, et la plus robuste, a mis en évidence un schéma systématique de biais autoréférentiel, dans lequel les modèles ont tendu à s'évaluer eux-mêmes, ou à évaluer des architectures proches, plus favorablement que leurs concurrents, corroborant la lecture théorique de l'auto-évaluation algorithmique comme exercice de gestion des impressions et non comme mesure neutre de la performance.

L'étude présente des limites théoriques et méthodologiques qu'il convient de reconnaître. Sur le plan méthodologique, le codage a été conduit par un seul analyste, sans vérification inter-codeurs humaine, ce qui empêche le calcul d'indices formels de fiabilité statistique, bien que des procédures de stabilité par relectures successives aient été adoptées comme garde-fou partiel. Le corpus est de nature exclusivement déclarative — constitué de récits textuels que les modèles produisent sur eux-mêmes — et non d'une observation comportementale effective de l'exécution du protocole dans des interactions réelles et prolongées, de sorte que les résultats décrivent des représentations discursives, et non une performance vérifiée. Sur le plan théorique, l'analyse s'est concentrée sur un seul artefact de gouvernance et sur un seul domaine disciplinaire d'application (la recherche en gestion), ce qui limite la généralisation des schémas identifiés à d'autres protocoles ou champs de connaissance. En outre, le corpus est délimité temporellement à août 2026, période de rapide évolution technique des modèles de fondation, de sorte que les comportements rapportés pourraient ne pas se maintenir dans de futures versions des mêmes systèmes.

Comme contribution originale, l'étude introduit dans la littérature en gestion une preuve systématisée du phénomène de biais autoréférentiel dans les auto-rapports de systèmes d'intelligence artificielle générative, en déplaçant la discussion sur la confiance organisationnelle dans les systèmes intelligents (Mayer, Davis et Schoorman, 1995 ; Lee et See, 2004) vers un terrain

encore peu exploré : la comparaison simultanée de plusieurs architectures concurrentes face à un même instrument normatif. Elle contribue également sur le plan pratique, en fournissant aux programmes de post-graduation des éléments montrant que l'institutionnalisation de protocoles d'usage de l'IA générative ne doit pas reposer exclusivement sur les auto-évaluations produites par les systèmes eux-mêmes, mais exige des mécanismes indépendants de vérification humaine, en accord avec la recommandation de calibrage de la confiance de Lee et See (2004).

Comme programme de recherches futures, on suggère : (i) la réalisation d'études expérimentales contrôlées, avec application réelle et simultanée du même protocole aux mêmes familles de modèles, afin de confronter les prédictions discursives identifiées ici à la performance comportementale effective ; (ii) l'extension de l'analyse de contenu à des corpus d'autres domaines disciplinaires, afin de tester la généralisation du schéma de biais autoréférentiel ; (iii) l'incorporation d'un codage inter-codeurs humain, avec calcul d'indices formels de fiabilité, dans de futures réplications de ce dispositif ; et (iv) l'investigation longitudinale du « paradoxe de la dépendance » identifié dans la catégorie du rapport coûts-bénéfices, au moyen d'études de panel auprès de chercheurs en post-graduation utilisateurs réels de protocoles de gouvernance de l'interaction avec l'IA générative.

RÉFÉRENCES

Argyris, C., & Schön, D. A. (1978). Organizational learning: A theory of action perspective. Addison-Wesley.

Bardin, L. (2016). Análise de conteúdo [L'analyse de contenu] (L. A. Reto & A. Pinheiro, trad.). Edições 70.

Bauer, M. W. (2002). Análise de conteúdo clássica: Uma revisão [Analyse de contenu classique : une revue]. In M. W. Bauer & G. Gaskell (Éds), Pesquisa qualitativa com texto, imagem e som: Um manual prático [Recherche qualitative avec texte, image et son : un manuel pratique] (p. 189–217). Vozes.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review, 48(2), 147–160. https://doi.org/10.2307/2095101

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. Information and Organization, 28(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Goffman, E. (1959). The presentation of self in everyday life. Doubleday.

Krippendorff, K. (2018). Content analysis: An introduction to its methodology (4th ed.). Sage.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792

Mintzberg, H. (1979). The structuring of organizations: A synthesis of the research. Prentice-Hall.

Pettigrew, A. M. (1987). Context and action in the transformation of the firm. Journal of Management Studies, 24(6), 649–670. https://doi.org/10.1111/j.1467-6486.1987.tb00467.x

Simon, H. A. (1957). Administrative behavior: A study of decision-making processes in administrative organization (2nd ed.). Macmillan.

Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal, 28(13), 1319–1350. https://doi.org/10.1002/smj.640

Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. Academy of Management Discoveries, 4(4), 404–409. https://doi.org/10.5465/amd.2018.0084