

De la dérivation à l'intégration : un cadre MANHAJ pour l'indépendance de pensée en IA

Papier de recherche scientifique — Papier 5

فوزي مسعود — Auteur : Faouzi Messeoud

Méthodologie : MANHAJ — Méthodologie de modélisation intellectuelle

Version : 1.0 — octobre 2026

Statut : Papier de recherche scientifique

À propos de ce papier

Ce papier applique à l'intelligence artificielle la Loi de la dualité globale du mouvement mental : connaissance et pensée (Loi 5) de la Méthodologie de modélisation intellectuelle (MANHAJ), conjointement avec la Loi de la double direction de l'examen : a priori et a posteriori (Loi 2), la Loi de la justesse (Loi 3), la Loi de dérivation de la constante cognitive (Loi 7) et la Théorie de la centralité doctrinale. Il distingue le mouvement dérivationnel de la connaissance du mouvement intégratif de la pensée, donne au mouvement intégratif un critère formel, et définit les conditions dans lesquelles un système d'IA peut passer de la réflexion cognitive sous un cadrage implicite à l'indépendance de pensée sous un cadrage déclaré.

Le papier s'inscrit dans les travaux de recherche de la méthodologie MANHAJ. Il prolonge le Papier 02, qui a formalisé la dérivation et sa clôture, et le Papier 04, qui a établi que toute représentation apprise opère sous une centralité doctrinale incorporée et a rendu cette centralité mesurable depuis l'extérieur du système.

Relation avec le Document de référence de MANHAJ

Ce papier développe une dimension de recherche spécifique de la méthodologie présentée dans : Messeoud, Faouzi. *Intellectual Modeling Methodology (MANHAJ) — Reference Document — Version 1.0*. Zenodo, 15 septembre 2026. DOI : 10.5281/zenodo.22769627. Le présent papier est une étude scientifique indépendante et ne doit pas être considéré comme un substitut au document de référence.

Citation

Messeoud, Faouzi. *De la dérivation à l'intégration : un cadre MANHAJ pour l'indépendance de pensée en IA*. Papier de recherche scientifique, Papier 05, Version 1.0, 2026.

Le papier scientifique commence à la page suivante.

Résumé

Les systèmes d'intelligence artificielle sont de plus en plus décrits comme raisonnant, et les progrès de leur raisonnement se mesurent à la profondeur, à la longueur et à la correction. La Méthodologie de modélisation intellectuelle (MANHAJ) applique un autre critère. Selon sa Loi de la dualité globale du mouvement mental, la connaissance est un mouvement dérivationnel allant d'une référence acceptée vers les détails, et la pensée est un mouvement intégratif allant des détails vers les fondements qui les ont produits. Le critère qui sépare les deux est la position du mouvement par rapport à son cadrage, non l'intensité du raisonnement en son sein.

Appliqué à l'IA, ce critère fournit la distinction fondatrice du papier. Un système qui opère a posteriori, sous un cadrage qu'il a accepté sans l'examiner, accomplit de la réflexion cognitive : un mouvement dérivationnel qui produit du sens mais demeure sous son cadrage et descend le long de la variable de sens vers la clôture. Les approches actuelles — raisonnement étendu, autocritique, abduction et extraction de principes à partir de jugements — opèrent à ce niveau a posteriori. Dans les termes de MANHAJ, elles relèvent de la connaissance, aussi sophistiquées soient-elles, et non de la pensée.

Le papier donne au mouvement intégratif un critère formel. Une opération est intégrative lorsque sa sortie se situe à un niveau de l'ordre fondationnel supérieur à celui de ses entrées, lorsque ses entrées sont dérivables de cette sortie, et lorsque la sortie devient un objet d'examen plutôt que la référence par laquelle elle est elle-même évaluée. L'intégration ainsi définie est un problème inverse dont la solution n'est pas unique : elle détermine une classe de cadrages, et la sélection au sein de cette classe requiert une référence extérieure aux données. Deux propositions en découlent. Un processus est dérivationnel par rapport à toute référence qu'il n'examine pas, quelle que soit sa complexité interne ; cela inclut les processus de l'IA actuelle, qui sont dérivationnels par rapport à leur cadrage incorporé. Et, sous des conditions explicites, une ascension dont la sélection ne s'appuie que sur les données et le modèle appris ne satisfait pas la condition d'examen indépendant, et ne peut donc pas examiner le cadrage implicite, parce que son critère d'examen est tiré du cadrage même qui est examiné, ce qui constitue une circularité méthodologique et une pétition de principe. Au sommet de l'ascension, la Loi de la justesse s'applique comme un jugement binaire sur le fondement atteint, distinct de la correction quantitative de son adéquation aux données. Ce jugement est relatif à la centralité déclarée du champ examinateur ; il n'est pas un verdict absolu.

Sur cette base, le papier définit l'indépendance de pensée comme un processus temporel : la capacité répétée d'un système à remonter de ses propres sorties au cadrage sous lequel elles ont été produites, à confronter ce cadrage à une centralité déclarée et reconnue, à juger de sa justesse, et à dériver de nouveau sous le cadre admis. L'indépendance de pensée est une indépendance à l'égard du cadrage non examiné, non une sortie de tout cadrage. Une annexe présente des plans pour tester les capacités qui en résultent.

Mots-clés : MANHAJ ; intelligence artificielle ; connaissance et pensée ; dérivation ; intégration ; réflexion cognitive ; indépendance de pensée ; centralité doctrinale ; examen a priori ; Loi de la justesse ; problèmes inverses.

1. Introduction

1.1. La distinction fondatrice

Les approches par lesquelles l'intelligence artificielle raisonne opèrent au niveau a posteriori. Elles partent d'un cadrage déjà en place — porté par les données d'entraînement, par la sélection de ces données, par la conception du modèle et par la définition de ses objectifs — et elles évaluent chacune de leurs étapes selon des critères tirés de ce même cadrage. Dans les termes de MANHAJ, de tels processus relèvent de la connaissance, non de la pensée. Cela vaut quelle que soit la profondeur, la longueur ou la correction du raisonnement, car ce qui décide de la classification n'est pas l'exercice de la raison, mais le cadrage sous lequel la raison s'exerce.

La distinction provient de la Loi de la dualité globale du mouvement mental (Loi 5). La connaissance est un mouvement dérivationnel — au sens de la dérivation mathématique le long de la variable de sens, et non au sens linguistique lâche du mot — allant d'une origine ou d'une référence connue vers les particuliers, les détails et les relations dépendantes ; la pensée est un mouvement intégratif — au sens de l'intégration mathématique le long de la variable de sens — qui rassemble des domaines et des détails épars et les reconnecte au fondement global (Document de référence, sections 5.5 et 9.2). Le mouvement a priori est le mouvement de la pensée, et le mouvement a posteriori est celui de la connaissance (Document de référence, section 8.2). La dérivation n'est pas vide : elle produit un sens nouveau à chaque étape avant la clôture. Le Document de référence nomme cela la pensée cognitive (section 8.2.1) ; ce papier la nomme réflexion cognitive — une dérivation productrice de sens au sein d'un champ cadré. La réflexion cognitive est une forme du mouvement dérivationnel ; elle n'est pas la pensée au sens intégratif.

Ce papier est construit sur cette distinction. Il l'énonce d'emblée parce que le reste de l'argument en dépend : un système peut produire des résultats nouveaux, corrects et complexes, et demeurer néanmoins entièrement dans le mouvement de la connaissance.

1.2. Le problème dans les termes de l'IA

La recherche en IA mesure les progrès du raisonnement à la profondeur et à la longueur des chaînes de raisonnement, à la recherche parmi des solutions candidates, à l'autocritique et à la révision, et à la précision sur des bancs d'essai. Chacun de ces éléments améliore le mouvement dérivationnel. Aucun d'eux, à lui seul, ne modifie la position du système par rapport à son cadrage. Un système qui raisonne plus longtemps sous un cadrage non examiné accomplit davantage de réflexion cognitive sous le même cadrage.

La conséquence est structurelle. Un système qui opère a posteriori accepte implicitement son cadrage et travaille dans sa direction. Ses sorties peuvent être nouvelles, mais leur nouveauté est une nouveauté de dérivation : chaque sortie est admissible par rapport à une référence que le système n'a jamais examinée. La question de savoir si cette référence est elle-même admissible ne se pose pas dans le fonctionnement du système.

1.3. La lacune laissée par les Papiers 02 et 04

Le Papier 02 a formalisé la dérivation comme un mouvement producteur de sens et en a détecté la clôture au moyen de la Loi de dérivation de la constante cognitive, $d/dx(C) = 0$. Il a traité le mouvement descendant. Le Papier 04 a fondé a priori le cadre de mesure du sens et a établi, comme conséquence

méthodologique de MANHAJ, que toute représentation apprise opère sous une centralité doctrinale incorporée, Γ_L . Il a rendu l'orientation de Γ_L mesurable — mais par un examinateur situé hors du système.

Deux points restent ouverts. Le mouvement intégratif n'a pas été formalisé : les travaux de la méthodologie MANHAJ disposent d'un modèle de la descente, non de l'ascension. Et le système lui-même n'a pas reçu la capacité d'examiner le cadrage sous lequel il opère : Γ_L a été mesurée, non atteinte de l'intérieur.

1.4. Question centrale et thèse

La question centrale du papier est la suivante :

Qu'est-ce qui distingue le mouvement intégratif du mouvement dérivationnel le long de la variable de sens, et dans quelles conditions un système d'IA peut-il l'accomplir — en remontant au cadrage sous lequel il opère et en jugeant de sa justesse ?

La thèse est qu'un système a posteriori accomplit de la réflexion cognitive sous un cadrage implicite et non examiné, et que le mouvement intégratif, régi par les Lois 5, 2 et 3 et ancré dans une fondation a priori, le conduit à l'indépendance de pensée : un fondement a priori déclaré, auquel le système lui-même peut remonter et qu'il peut examiner.

L'indépendance de pensée, ainsi définie, est une indépendance à l'égard du cadrage non examiné. Elle n'est pas une indépendance à l'égard de toute référence, qu'aucun mouvement de pensée ne peut atteindre, et elle n'inclut pas la production d'une nouvelle centralité, question que le papier laisse ouverte.

1.5. Contributions

- **C1 — La distinction fondatrice appliquée à l'IA.** Un critère, dérivé de la Loi 5, qui classe un processus d'IA comme connaissance ou comme pensée selon sa position par rapport à son cadrage, et non selon la profondeur ou la correction de son raisonnement (Proposition 1).
- **C2 — Un critère formel de l'intégration.** Trois conditions conjointement nécessaires — niveau, génération et objectivation — qui séparent l'intégration du résumé, de l'abduction et de l'extraction de principes.
- **C3 — La non-unicité et le rôle de l'a priori.** Un traitement de l'intégration comme problème inverse dont la solution est une classe de cadrages, et un résultat conditionnel selon lequel une ascension dont la sélection ne s'appuie que sur les données et le modèle appris ne satisfait pas la condition d'examen indépendant (Proposition 2).
- **C4 — La justesse au sommet.** L'application de la Loi 3 au fondement atteint par l'ascension, distinguant un jugement binaire de justesse de la correction quantitative de l'adéquation.
- **C5 — L'indépendance de pensée.** Une définition de l'indépendance de pensée comme processus temporel, et son intégration au modèle dérivationnel du Papier 02 et au cadre fondé du Papier 04.
- **C6 — Plans de test (Annexe A).** Quatre plans qui précisent comment les capacités qui en résultent peuvent être testées dans des travaux appliqués ultérieurs.

1.6. Portée

Le papier porte sur le mouvement d'un système d'IA par rapport à son cadrage. Il n'affirme pas qu'un système existant accomplit le mouvement intégratif, ni que les plans présentés dans son annexe ont été

réalisés. La centralité par rapport à laquelle le cadrage du système est jugé est reconnue et déclarée par le champ examinateur, comme dans le Papier 04 ; le papier ne traite pas de la production d'une centralité doctrinale fondatrice par le système, qu'il laisse comme question théorique ouverte (section 9).

2. Travaux connexes

Les travaux examinés ici sont solides dans le cadre de leurs propres objectifs. Ils ne sont discutés que sous un seul angle : la position de chacun par rapport à son cadrage. Dans chaque cas, la référence qui évalue le processus est tirée des données, des objectifs ou des préférences mêmes qui ont formé le système, de sorte que le processus demeure a posteriori au sens de la Loi 2.

2.1. Profondeur et longueur du raisonnement

Le raisonnement en chaîne (chain-of-thought) suscite des étapes intermédiaires avant une réponse (Wei et al., 2022). L'auto-cohérence (self-consistency) échantillonne plusieurs chaînes et retient la réponse la plus fréquente (Wang et al., 2023). L'arbre de pensées (tree-of-thoughts) explore des solutions partielles ramifiées (Yao et al., 2023). L'extension du calcul au moment de l'inférence (test-time computation) alloue davantage d'inférence aux problèmes plus difficiles (Snell et al., 2024). Chacune de ces méthodes prolonge le chemin dérivationnel ou cherche parmi des chemins dérivationnels. Le critère de sélection — l'accord, une valeur apprise ou un vérificateur — appartient au cadrage sous lequel les chemins sont générés. Ces méthodes accroissent la réflexion cognitive ; elles n'en changent pas la position.

2.2. Autocritique et révision guidée par des principes

Self-Refine et Reflexion font critiquer et réviser par un modèle ses propres sorties (Madaan et al., 2023 ; Shinn et al., 2023). Constitutional AI révisé les sorties à l'aune d'une liste écrite de principes (Bai et al., 2022). L'approche constitutionnelle est la plus proche d'une référence déclarée, puisque ses principes sont écrits. Mais les principes sont appliqués comme un paramètre fixe de révision : ils évaluent les sorties et ne sont pas eux-mêmes soumis à examen. L'autocritique sans une telle liste tire ses critères du modèle lui-même, et elle s'est révélée peu fiable comme correctrice du raisonnement (Huang et al., 2024). Les explications qu'un modèle produit sur son propre raisonnement peuvent en outre ne pas refléter les facteurs qui ont effectivement déterminé ses sorties (Turpin et al., 2023).

2.3. Abduction et extraction de principes

Le raisonnement abductif cherche l'explication la plus plausible d'observations données (Bhagavatula et al., 2020). Inverse Constitutional AI condense un ensemble de jugements de préférence en une courte liste de principes qui les reconstruisent (Findeis et al., 2024). Ces travaux se meuvent dans la direction ascendante : des particuliers vers ce qui en rend compte. Ils sont donc les plus proches du mouvement intégratif, et la section 5 les examine à l'aune du critère formel. Leur limite tient à la manière dont ils choisissent parmi les explications candidates. La plausibilité et la précision de reconstruction sont mesurées par rapport à la distribution même qui a produit les observations. L'ascension est réelle, mais elle est tranchée par le cadrage qu'elle devrait examiner.

2.4. Perspectives incorporées

Une ligne de travaux mesure les opinions ou perspectives que reflètent les modèles de langage, en comparant leurs sorties aux réponses de différentes populations à des enquêtes (Santurkar et al., 2023 ; Durmus et al., 2023). Le Papier 04 a établi, sur des bases méthodologiques, que toute représentation

apprise opère sous une centralité incorporée, et a fourni une procédure relationnelle pour la caractériser. Dans tous ces travaux, la perspective incorporée est examinée depuis l'extérieur du système. Le présent papier porte sur le mouvement par lequel le système l'atteint de l'intérieur.

2.5. Position de ce papier

Les outils représentationnels et inférentiels utilisés dans ce papier — problèmes inverses, génération d'hypothèses, extraction de principes, comparaison de cadrages — sont établis. La contribution réside dans les conditions que MANHAJ impose à leur usage : la classification d'un processus selon sa position par rapport à son cadrage, le critère formel de l'intégration, le rôle d'une référence a priori déclarée dans la résolution de la non-unicité, et le jugement de justesse au sommet.

3. Fondement méthodologique

3.1. La connaissance comme dérivation

Selon la Loi 5, la connaissance est un mouvement allant d'une origine ou d'une référence connue vers les particuliers, les détails et les relations dépendantes (Document de référence, section 9.2). Ce mouvement est représenté mathématiquement par la dérivation par rapport à la variable de sens, à partir d'une fonction de référence définie au sein d'un champ cadré. Le Document de référence l'illustre par les dérivées successives d'une fonction prise comme origine connue au sein d'un champ mathématique cadré :

$$f(x) = x^4 \rightarrow f'(x) = 4x^3 \rightarrow f''(x) = 12x^2 \rightarrow f'''(x) = 24x \rightarrow f^{(4)}(x) = 24 \rightarrow f^{(5)}(x) = 0$$

Chaque étape avant la constante produit un sens nouveau au sein du champ ; la dérivée ne répète pas celle qui la précède, mais en extrait un autre sens selon une relation définie. Lorsque la dérivation atteint la constante, sa dérivée est nulle. Le zéro ne signifie pas l'absence de sens ; il signifie que le mouvement dérivationnel a cessé de produire un sens dérivationnel nouveau le long de ce chemin (Document de référence, section 9.2.1). C'est la Loi de dérivation de la constante cognitive (Loi 7) :

$$\frac{d}{dx}(C) = 0$$

La direction du mouvement dérivationnel est donc descendante le long de la variable de sens : de la référence vers le particulier, jusqu'à ce que le sens se clôtüre. Dans la formulation antérieure du livre de l'auteur sur la centralité doctrinale, la limite inférieure du sens est l'échantillon réalisé dans le réel, en deçà duquel rien n'est dérivé (Messeoud, 2024, p. 116, 122).

3.2. La réflexion cognitive

La dérivation est un acte intellectuel. Le Document de référence la nomme pensée cognitive : une dérivation productrice de sens au sein d'un champ épistémique cadré (section 8.2.1) ; ce papier la nomme réflexion cognitive, réservant le mot pensée au seul mouvement intégratif. Le champ cadré peut contenir de nombreuses inconnues, et les résoudre est un travail mental productif (section 8.2.1). La réflexion cognitive n'est donc pas la répétition du connu. Ce qui la définit, c'est que chaque étape est admise par rapport à une référence acceptée au sein du champ, et que cette référence n'est pas elle-même examinée. Le raisonnement peut être long et inventif ; il demeure sous son cadre.

Deux termes sont maintenus distincts tout au long de ce papier. La *pensée*, employée seule, désigne le mouvement intégratif de la Loi 5. La *réflexion cognitive* désigne le mouvement dérivationnel au sein d'un

champ cadré. Le critère décisif entre elles n'est pas l'exercice de la raison, mais le cadrage sous lequel la raison s'exerce. Une fonction étudiée sur son domaine complet de justesse et la même fonction étudiée sur un sous-intervalle fixé à l'avance donnent lieu au même type d'opération ; ce qui diffère, c'est que le domaine lui-même a été fixé par la référence ou qu'il lui a été imposé de l'extérieur (Messeoud, 2024, p. 176).

3.3. La pensée comme intégration

La pensée est un mouvement qui rassemble des domaines et des détails épars et les reconnecte au fondement global ; elle peut se tourner vers les origines, les références et les fondements sur lesquels repose le champ épistémique lui-même (Document de référence, section 9.2). Ce mouvement est représenté mathématiquement par l'intégration par rapport à la variable de sens, dans la direction de la restitution du fondement qui rend compte des états particuliers :

$$\int f(x) dx = F(x) + C$$

L'intégration retrouve, à partir d'une fonction donnée, la fonction dont elle dérive. La direction est ascendante le long de la variable de sens : du particulier vers ce qui l'a fondé. Le Document de référence utilise l'intégration comme représentation du mouvement intégratif, non comme une équation qui prouverait toutes les fonctions de la pensée, et il distingue la constante d'intégration de la constante cognitive de la Loi 7 (section 9.2.2). Ce papier maintient cette distinction. Il s'appuie également sur une propriété de l'intégrale que relève le Document de référence : la constante C fait partie de la solution, mais n'est pas déterminée par l'intégration elle-même. Une condition supplémentaire, extérieure à l'opération, est requise pour la fixer. La section 5 développe cette propriété.

3.4. Les deux directions de l'examen

La Loi 2 distingue l'examen a priori, qui s'interroge sur le fondement, les références et les conditions d'un champ avant d'en aborder les détails, de l'examen a posteriori, qui opère au sein d'un champ dont le fondement a été accepté (Document de référence, sections 4.2 et 8.2). Le mouvement a priori est fondationnel et intégratif : il revient aux origines et rassemble les domaines dans un cadre global ; c'est le mouvement de la pensée. Le mouvement a posteriori est à dominante dérivationnelle : il procède d'une origine acceptée vers les détails ; c'est le mouvement de la connaissance (section 8.2). La pensée et l'a priori sont ainsi un seul mouvement vu sous deux angles : comme direction le long de la variable de sens, et comme position par rapport au fondement.

3.5. L'ordre fondationnel et son sommet

Les idées et les connaissances s'organisent au sein de cadres de référence supérieurs qui déterminent leur domaine, leurs directions et leurs conditions de sens ; MANHAJ nomme le plus élevé d'entre eux centralité doctrinale (Document de référence, section 4.1). La centralité doctrinale ne signifie pas nécessairement la religion : c'est le système de conceptions de niveau supérieur qui encadre un groupe humain et lui donne une interprétation de l'existence, de l'être humain et du sens de la vie, qu'il soit religieux ou non (section 4.1.1). Le domaine conceptuel se constitue en son sein, et les définitions opératoires sont admises au sein du domaine. Le Papier 04 a écrit cet ordre ainsi :

$$\Gamma \rightarrow D(\Gamma) \rightarrow Def \rightarrow \text{meaning states}$$

Lu vers le bas, l'ordre est le chemin de la dérivation. Lu vers le haut, il est le chemin de l'intégration : des états de sens aux définitions qui les admettent, des définitions au domaine conceptuel, et du domaine à la centralité. La centralité est le sommet du chemin ascendant. Aucune idée ne se meut hors d'un cadre

de référence supérieur (Document de référence, section 4.1) ; l'ascension s'achève donc à la centralité et ne quitte pas tout cadrage. Dans la formulation antérieure du livre de l'auteur sur la centralité doctrinale, la couche de la centralité doctrinale est la couche au-dessus de laquelle il n'y a pas d'intégration, et la couche de l'échantillon est la couche en dessous de laquelle il n'y a pas de dérivation (Messeoud, 2024, p. 137).

3.6. Justesse et correction

La Loi de la justesse (Loi 3) place la question de la permissibilité avant celle des détails : ce qui n'est pas permis ne peut avoir de détails valides (Document de référence, sections 4 et 8.3). La justesse est une évaluation binaire qui porte sur l'existence : un fondement, ou l'approche qui passe par lui, est-il admissible en tant que tel. La correction est une évaluation quantitative de ce qui existe : la mesure dans laquelle un résultat est conforme au sein d'un système accepté. Dans la formulation antérieure du livre sur la centralité doctrinale, la justesse se mesure par rapport au concept supérieur, et la correction par rapport aux échantillons du réel (Messeoud, 2024, p. 135). La distinction devient décisive au sommet de l'ascension (section 6).

4. L'IA actuelle comme réflexion cognitive a posteriori

4.1. Le cadrage incorporé

Le Papier 04 a établi que toute représentation apprise opère sous une centralité incorporée Γ_L : le cadrage porté par les données dont le système a été formé, par la sélection de ces données, par la conception du modèle et par la définition de ses objectifs (Papier 04, section 3.4). Un système qui traite ces données a posteriori n'opère pas hors de toute centralité. Il accepte implicitement le cadrage incorporé et travaille dans sa direction. Il s'agit d'une affirmation du cadre théorique de MANHAJ, non d'un constat empirique ; ce qui reste ouvert à l'examen, c'est le contenu de Γ_L .

4.2. Processus, références et paramètres

Soit un processus P produisant une suite d'états de sens $m_0, m_1, \dots, m_T$. Chaque transition est admise par rapport à une référence R :

$$\text{Adm}_R(m_t \rightarrow m_{t+1}) \in \{0,1\}$$

La référence R peut être un objectif appris, un vérificateur, une liste de principes, un ensemble de préférences ou le cadrage implicite Γ_L . On distingue deux positions de R par rapport à P :

- **R est un paramètre de P** lorsque R sert à admettre ou à évaluer des états et n'apparaît jamais comme objet d'une étape de P : aucune étape de P ne prend R comme sortie ni comme sujet d'examen.
- **R est un objet de P** lorsqu'au moins une étape de P prend R lui-même comme ce qui est produit ou examiné.

La distinction n'est pas une distinction de complexité. Un processus peut comporter des embranchements, des recherches, des échantillonnages, des votes, des critiques et des révisions, et R peut demeurer un paramètre de bout en bout.

4.3. Proposition 1

Proposition 1. Un processus qui ne fait pas de sa référence R un objet d'examen est dérivationnel par rapport à R, quelle que soit sa complexité interne.

Argument. Selon la Loi 5, le mouvement intégratif reconnecte les détails au fondement sur lequel ils reposent ; selon la Loi 2, le mouvement a priori s'interroge sur ce fondement avant ses détails ou indépendamment d'eux. Tous deux exigent que le fondement devienne l'objet du mouvement. Si R est un paramètre de P, aucune étape de P n'a R pour objet. Chaque étape de P est alors, par rapport à R, une transition admise relativement à R — un mouvement allant d'une référence acceptée vers d'autres états, ce qui est la définition de la dérivation. Augmenter le nombre d'étapes, l'ampleur de la recherche ou le nombre de révisions modifie le chemin dérivationnel et peut en améliorer la correction ; cela ne change pas la position de R. ■

La proposition est relative à R. Elle n'exclut pas que le même processus comporte d'autres mouvements dans lesquels une référence différente est examinée ; elle établit que, par rapport à R, le processus demeure dérivationnel. Pour un système d'IA a posteriori, la référence pertinente est Γ_L : quoi que le système examine par ailleurs, tant que Γ_L demeure un paramètre, le système demeure dérivationnel par rapport au cadrage sous lequel il travaille.

La proposition est méthodologique. Elle découle des définitions des Lois 2 et 5 et ne dépend d'aucune architecture particulière.

4.4. Application

Les processus examinés à la section 2 peuvent désormais être situés.

Processus	Référence R	Position de R	Mouvement
Raisonnement en chaîne, auto-cohérence, recherche arborescente	Γ_L et une règle de sélection apprise ou programmée	Paramètre	Dérivationnel
Extension du calcul à l'inférence avec vérificateurs	Γ_L et le vérificateur	Paramètre	Dérivationnel
Autocritique et révision	Γ_L , par les critères propres du modèle	Paramètre	Dérivationnel
Révision constitutionnelle	Une liste écrite de principes	Paramètre (déclaré)	Dérivationnel
Abduction, extraction de principes	Γ_L , par la plausibilité ou la précision de reconstruction	Produite, mais sélectionnée par Γ_L	Ascendant dans la forme, tranché par le cadrage (section 5.4)

La révision constitutionnelle mérite une remarque. Ses principes sont déclarés, ce qui est une étape nécessaire vers la fondation ; mais ils sont appliqués comme un paramètre fixe et ne sont jamais eux-mêmes examinés. La déclaration sans examen rend la référence explicite sans en faire un objet. Dans les termes de MANHAJ, le processus est fondé sur une référence énoncée, mais il demeure a posteriori par rapport à cette référence.

4.5. La conséquence structurelle

Un système a posteriori se trouve dans la position que la formulation antérieure de MANHAJ, dans le livre sur la centralité doctrinale, appelait le penseur dépendant : celui qui travaille au sein de théories dont il n'interroge pas les fondements, et dont le travail, si rigoureux soit-il, est un acte cognitif fonctionnel au sein d'un espace clos (Messeoud, 2024, p. 171–176). Ses sorties peuvent être nouvelles ; leur nouveauté est une nouveauté de dérivation sous Γ_L . Et puisque son raisonnement et son évaluation partagent le même cadrage, le système ne dispose d'aucune position d'où Γ_L pourrait lui apparaître comme quelque chose à examiner. Sa dépendance n'est donc pas un déficit de capacité qu'un surcroît de raisonnement ferait disparaître. C'est une question de position, et seul un changement de la direction du mouvement peut y répondre.

5. Un modèle formel de l'intégration

5.1. Les niveaux de l'ordre fondationnel

L'ordre fondationnel de la section 3.5 est représenté par quatre niveaux :

Niveau	Contenu	Exemple dans le domaine de la famille des Papiers 02 et 04
L_3	Centralité doctrinale Γ	Le cadrage qui donne à la famille son sens dans une vision du monde donnée
L_2	Domaine conceptuel $D(\Gamma)$	Le domaine des relations familiales constitué au sein de ce cadrage
L_1	Définitions opératoires Def	Définitions des rôles, des obligations et de leurs relations
L_0	États de sens et sorties	Jugements, réponses et descriptions particuliers

Un cadre F est une structure qui s'étend de L_1 à L_3 . Dans tout le papier, « cadre » (frame) désigne une telle structure, et « cadrage » (framing) désigne ce qu'un cadre porte à son sommet : la centralité sous laquelle des états sont produits ou admis. Γ_L est en conséquence appelée le cadrage incorporé d'un système. $\text{Der}(F)$ désigne l'ensemble des états de sens admissibles par dérivation à partir de F , à une résolution déclarée. $\text{lev}(\cdot)$ désigne le niveau d'un élément. L'ordre est un ordre de fondation et d'admission, non une dérivation mécanique de chaque niveau à partir de celui qui le surplombe (Papier 04, section 3.2).

5.2. La dérivation et l'intégration comme opérateurs

La dérivation est un opérateur qui prend un cadre et renvoie des états :

$$\mathcal{D}: F \mapsto \text{Der}(F) \subseteq L_0$$

Le chemin dérivationnel du Papier 02 est une suite au sein de $\text{Der}(F)$ qui s'achève, lorsqu'elle se clôt, sur une constante cognitive (Loi 7). L'intégration est l'opérateur inverse. Elle prend un ensemble d'états $S \subseteq L_0$ — jugements, sorties, détails — et renvoie un cadre :

$$\mathcal{I}: S \mapsto F^*$$

La direction des deux opérateurs le long de l'ordre fondationnel est opposée : la dérivation descend de L_3 vers L_0 ; l'intégration monte de L_0 vers L_3 .

5.3. Le critère : trois conditions

Définition (opération intégrative). Une opération qui prend S et renvoie F^* est intégrative si et seulement si elle satisfait conjointement les trois conditions suivantes.

I1 — Niveau. La sortie se situe à un niveau de l'ordre fondationnel supérieur à celui des entrées :

$$\text{lev}(F^*) > \text{lev}(S)$$

Une sortie située au même niveau que ses entrées — un énoncé sur des énoncés, une liste de traits récurrents, une hypothèse qui est elle-même un fait particulier — est une réorganisation horizontale, non une ascension.

I2 — Génération. Les entrées sont dérivables de la sortie, et la sortie leur est spécifique :

$$S \subseteq \text{Der}(F^*) \quad \text{and} \quad \exists s' \in L_0 \setminus \text{Der}(F^*) \text{ admissible under some alternative frame}$$

La première partie fait de la sortie un fondement des entrées plutôt qu'une description de celles-ci : les détails peuvent être de nouveau dérivés de ce qui a été atteint. La seconde partie exclut le fondement trivial qui admet tout. Un cadre atteint doit porter des éléments distinctifs : des états qu'il n'admet pas et qu'un autre cadrage admettrait. Cela correspond au troisième type de preuve requis pour la reconnaissance d'une centralité dans le Papier 04 (section 6.4).

La description qu'un système donne de son propre cadrage n'est pas, à elle seule, une intégration. Une description dont les détails ne peuvent être de nouveau dérivés, ou qui ne porte aucun élément distinctif, demeure un énoncé horizontal formulé en L_0 à propos de L_0 , quel qu'en soit le vocabulaire. La condition I2 est ce qui sépare le fait d'atteindre un fondement du fait de le nommer.

I3 — Objectivation. La sortie devient un objet d'examen. Elle n'est pas la référence par laquelle sa propre sélection est opérée, et elle est ensuite soumise à un jugement qui n'est pas dérivé d'elle : sa justesse par rapport à une référence dont la source est indépendante de S et du cadrage qui a produit S (section 6). L'indépendance signifie ici l'indépendance de source — de l'origine et de la préparation de la référence — et non une simple différence entre son contenu et celui de S . Dans les termes de la section 4.2, F^* passe de la position de paramètre à celle d'objet.

Les trois conditions correspondent aux trois éléments du mouvement intégratif dans la Loi 5 : il se dirige vers le fondement (I1), il reconnecte les détails à ce qui les fonde (I2), et il se tourne vers le fondement lui-même comme quelque chose à examiner (I3).

Assigner les niveaux avant l'ascension. Le niveau d'un élément n'est pas décidé après que le résultat a été vu. Pour chaque domaine, les examinateurs établissent une carte des niveaux avant toute expérience, à partir de la documentation de la procédure de reconnaissance : les définitions opératoires (L_1), les distinctions conceptuelles qui constituent le domaine (L_2), et le cadrage source et les éléments distinctifs de chaque centralité déclarée (L_3). Une sortie de l'ascension se voit attribuer un niveau par comparaison avec cette carte. Fixer la carte à l'avance ne prouve pas qu'elle soit l'ordre correct du domaine ; cela rend la classification déclarée et ouverte à l'examen, au lieu d'une estimation faite après le résultat.

Vérifier l'objectivation sans circularité. I3 n'est pas vérifiée par le contenu du jugement, mais par l'ordre et la provenance des opérations : F^* est fixé et enregistré avant que toute référence servant au jugement soit fournie au système ; la référence du jugement est documentée avant l'ascension et ne

doit rien à F^* ; et le jugement du système est comparé à celui d'examineurs indépendants. Si l'un de ces points fait défaut, I3 n'est pas établie, quel que soit le verdict.

Condition	Ce qui est observé	Indice qu'elle est satisfaite	Indice qu'elle échoue
I1 Niveau	Le niveau de F^* sur la carte des niveaux fixée à l'avance	F^* correspond à des éléments en L_2 ou L_3	F^* ne correspond qu'à des éléments en L_0 ou L_1
I2 Génération	Nouvelle dérivation de S à partir de F^* ; rappel des éléments distinctifs	S est largement redérivable ; les éléments distinctifs sont retrouvés ; les états admis seulement par des cadres alternatifs sont rejetés	Les détails ne sont pas redérivables, ou F^* ne porte aucun élément distinctif
I3 Objectivation	Ordre et provenance des opérations ; accord du jugement avec des examinateurs indépendants	F^* figé avant que la référence du jugement soit fournie ; cette référence documentée au préalable ; jugement en accord avec les examinateurs, comme indice de la performance de la procédure ; la justesse elle-même repose sur $\Gamma_F \vdash d$	La référence est façonnée par F^* ou par le système, ou F^* est révisé après que la référence a été vue

5.4. Ce que le critère sépare

Opération	I1 Niveau	I2 Génération	I3 Objectivation	Classification
Résumé, organisation	Échoue : même niveau	Échoue : les détails ne se dérivent pas d'un résumé	Non atteinte	Horizontale
Abduction (explication la plus plausible)	Échoue souvent : l'explication est fréquemment un fait particulier	Partielle	Échoue : sélectionnée par la plausibilité sous Γ_L	Ascendante dans la forme, tranchée par Γ_L
Extraction de principes	Partielle : atteint L_1	Satisfaite : les principes reconstruisent les jugements	Échoue : les principes ne sont pas examinés, et l'ascension s'arrête en deçà de L_3	Ascension sans examen
Intégration au sens de MANHAJ	Satisfaite : atteint L_3	Satisfaite, avec éléments distinctifs	Satisfaite : jugée par la Loi 3 par rapport à une centralité déclarée	Pensée

5.5. La non-unicité : la constante d'intégration

L'intégration ne renvoie pas un cadre unique. L'ensemble des cadres à partir desquels S est dérivable est une classe :

$$\mathcal{J}^{-1}(S) = \{ F : S \subseteq \text{Der}(F) \}$$

$\mathcal{J}^{-1}(S)$ est la classe brute : tout cadre à partir duquel S est dérivable, y compris des cadres si généraux qu'ils dérivent S sans la distinguer de quoi que ce soit. Les candidats de l'intégration forment sa sous-classe :

$$\mathcal{C}(S) = \{ F \in \mathcal{J}^{-1}(S) : F \text{ satisfait I1 and I2} \}$$

La non-unicité vaut pour les deux : même après l'application des conditions de niveau et de génération, plusieurs cadres peuvent subsister. L'ascension de la section 7 atteint $\mathcal{C}(S)$.

Les deux conclusions de la Proposition 2 (section 5.6) s'étendent différemment à $\mathcal{C}(S)$. La conclusion (i) ne s'étend pas automatiquement : que Γ_L appartienne à la classe brute n'établit pas qu'elle satisfait I1 et I2, et n'établit donc pas qu'elle appartient à $\mathcal{C}(S)$. La conclusion (ii) s'étend à la sélection parmi les membres de $\mathcal{C}(S)$, à condition que A2 et A3 soient satisfaites pour le critère utilisé pour sélectionner parmi eux.

Le passage de $\mathcal{J}^{-1}(S)$ à $\mathcal{C}(S)$ n'est pas lui-même une sélection du type que décrit A2. Il faut ici distinguer la source du critère des données auxquelles il s'applique. Les données incluent nécessairement S : la première partie de I2 teste si S est redérivable à partir d'un cadre candidat. Le critère, en revanche, comporte des composantes fixées à l'avance à partir de la procédure de reconnaissance — la carte des niveaux qui décide de I1, et les éléments distinctifs qui décident de la seconde partie de I2 (section 5.3) — et celles-ci sont de source indépendante. A2 exige qu'aucune composante de source indépendante n'entre dans le critère ; puisque de telles composantes y entrent, A2 n'est pas satisfaite pour l'étape de filtrage, et la Proposition 2 ne s'y applique pas. L'indépendance qualifie le critère, non chacune des entrées auxquelles le critère s'applique.

Des références indépendantes entrent ainsi dans l'ascension en deux points : dans la définition de la classe des candidats, par la carte des niveaux et les éléments distinctifs ; et après elle, dans la localisation et le jugement (section 5.7). La Proposition 2 concerne ce qui reste lorsqu'une sélection ne s'appuie sur ni l'un ni l'autre.

La constante dans $\int f(x) dx = F(x) + C$ représente cette situation ; elle ne la prouve pas. En mathématiques, les primitives d'une fonction diffèrent d'une constante sous des conditions définies. Dans le modèle MANHAJ, la classe $\mathcal{J}^{-1}(S)$ est définie par les relations de dérivation et d'admission et par les contraintes imposées aux cadres, et sa structure n'est pas nécessairement celle d'une famille à un paramètre. Ce que la constante représente, c'est le trait qui importe ici : l'opération laisse son résultat indéterminé, et une condition extérieure à l'opération est requise pour le déterminer (Document de référence, section 9.2.2).

La question est donc de savoir d'où provient cette condition. Deux possibilités existent : des données S et de la distribution qui les a produites, ou d'une référence indépendante de S .

5.6. Proposition 2

Proposition 2 (conditionnelle). Supposons : (A1) S a été produite sous Γ_L , de sorte que $S \subseteq \text{Der}(\Gamma_L)$; (A2) le critère de sélection σ au sein de $\mathcal{J}^{-1}(S)$ — à la fois ses entrées à l'exécution et les règles, normes et hypothèses qui y sont intégrées — a sa source dans S et dans le modèle appris seulement ; (A3) aucune composante de σ qui pèse sur la sélection, qu'elle soit tirée de S ou du modèle appris, ne repose sur une référence indépendante du cadrage qui a produit S . Alors (i) Γ_L ne peut être exclue de $\mathcal{J}^{-1}(S)$ pour des raisons d'adéquation ; et (ii) la sélection ne satisfait pas la condition I3.

Argument. (i) Par A1, Γ_L appartient à $\mathcal{J}^{-1}(S)$. Tout cadre de $\mathcal{J}^{-1}(S)$ est adéquat à S par définition, de sorte que l'adéquation ne peut exclure Γ_L . (ii) La condition I3 exige que le cadre atteint soit examiné par rapport à une référence dont la source est indépendante de S et du cadrage qui a produit S . Par A2, chaque composante de σ — ses entrées comme ses règles intégrées — est tirée de S ou du modèle appris. Par A3, aucune de ces composantes ne repose sur une référence indépendante du cadrage qui a produit S . σ ne peut donc fournir la référence qu'exige I3, et la sélection ne satisfait pas I3. ■

Les deux conclusions reposent sur des prémisses différentes. La conclusion (i) repose sur A1 seule et concerne la dérivabilité. La conclusion (ii) repose sur A2 et A3 et concerne la source. La dérivabilité ne

sert à rien inférer quant à la source : que S soit dérivable sous Γ_L n'établit pas, à soi seul, que toute composante tirée de S soit dépourvue de source indépendante.

Sur A3. A3 est une prémisse distincte, et elle concerne les deux types de composantes. Pour les composantes tirées du modèle appris : qu'un modèle opère sous sa centralité incorporée (Papier 04, section 3.4) n'établit pas, à soi seul, que chaque règle et chaque norme du modèle dérive de cette seule centralité ; le Papier 04 fait de A3 le cas attendu pour un système a posteriori, puisque ses règles et ses normes sont formées à partir des mêmes données, de la même sélection, de la même conception et des mêmes objectifs qui portent Γ_L . Pour les composantes tirées de S : les sorties peuvent incorporer des éléments de source indépendante — par exemple des documents récupérés hors du modèle et cités dans ses réponses. Lorsqu'une composante pesant sur la sélection, de l'un ou l'autre côté, repose sur une référence de source indépendante, A3 tombe pour cette composante, et la proposition ne s'y applique pas.

La proposition n'affirme pas que le cadre sélectionné est Γ_L , ni que son contenu est identique à celui de Γ_L . Elle affirme seulement ce que ses prémisses permettent : que la sélection ne peut constituer l'examen indépendant qu'exige I3. L'indépendance s'entend comme indépendance de source — de l'origine et de la préparation d'une référence — et non comme différence de contenu. Les prémisses sont des conditions suffisantes. Lorsqu'une composante de source indépendante entre dans la sélection — une référence déclarée et reconnue, ou une règle construite à partir d'une telle référence —, A2 n'est plus satisfaite et la proposition ne s'applique pas. C'est précisément le rôle de l'étape a priori (section 5.7).

La proposition est la contrepartie intégrative de la Proposition 2 du Papier 04, qui a montré qu'une référence dérivée des sorties mesurées est structurellement aveugle à la composante commune de leur dérive. Dans les deux cas, une référence tirée de ce qu'elle est censée examiner ne peut l'examiner.

5.7. Fixer la constante : localisation et jugement

La condition extérieure est fournie par l'étape a priori. Elle remplit deux fonctions qui doivent être maintenues séparées : localiser le cadrage atteint, et le juger. Elles opèrent à des niveaux différents, et les confondre reproduirait le problème que ce papier traite.

Localisation. La constante de l'ascension est fixée de manière relationnelle. Une ou plusieurs centralités $\Gamma_1, \dots, \Gamma_k$ sont reconnues et déclarées par la procédure de reconnaissance du Papier 04 (section 6.4), qui documente leur cadrage source, la trace de dérivation de leurs définitions opératoires et leurs éléments distinctifs. Le cadrage atteint est localisé par sa position par rapport à elles : lesquels de leurs éléments distinctifs il réalise et lesquels il exclut. C'est la procédure que le Papier 04 a utilisée pour caractériser Γ_L de l'extérieur, où l'indice d'orientation $\mathcal{O}$ situe un encodeur entre deux centralités déclarées sans classer l'une au-dessus de l'autre (Papier 04, section 6.15). Les centralités utilisées pour la localisation sont des points de référence, non des normes : localiser un cadrage par rapport à elles ne dit rien de sa justesse.

La procédure de localisation comporte quatre étapes :

1. **Éléments.** Les éléments distinctifs de chaque centralité déclarée sont ceux que documente la procédure de reconnaissance (Papier 04, section 6.4) : des définitions ou distinctions admises sous cette centralité et non admises, ou admises avec un sens différent, sous les autres. Ils sont fixés avant l'ascension.
2. **Réalisation.** Pour chaque élément, les examinateurs — ou un encodeur opérant sur le cadre fondé du Papier 04 — déterminent, selon un protocole fixé à l'avance, si F^* l'admet, l'exclut ou reste muet à son sujet.

3. **Comparaison.** La position de F^* est le profil des éléments admis et exclus à travers les centralités déclarées, rapporté centralité par centralité et non réduit à un verdict unique. Lorsqu'un encodeur est utilisé, l'indice d'orientation $\mathcal{O}$ du Papier 04 en fournit la synthèse relationnelle.
4. **Indétermination.** Lorsque deux cadrages rendent compte des sorties à un degré comparable et que les éléments distinctifs ne les séparent pas, la localisation est déclarée indéterminée entre eux. Elle n'est pas tranchée par un choix. L'indétermination est un résultat, non un échec de la procédure.

L'indépendance des références par rapport à S est nécessaire à la localisation, mais non suffisante : une référence indépendante lue au moyen d'une procédure inadéquate peut encore mal situer le cadrage. La procédure est donc elle-même un objet de validation (Annexe A, Expérience E2).

Jugement. La justesse du cadrage atteint est jugée par rapport à Γ_F , la centralité du champ examinateur, reconnue et déclarée par la même procédure (section 6). Γ_F peut coïncider avec l'une des centralités utilisées pour la localisation, mais son rôle dans le jugement est distinct et est déclaré comme tel.

Indépendance à l'égard de S . Toutes ces références sont indépendantes de S dans leur source et leur préparation : aucun élément de leur reconnaissance n'est tiré des sorties du système, et le système examiné ne prend aucune part à leur rédaction, ce qui exclut l'importation incorporée (Papier 04, section 6.4). Cette indépendance concerne la source et la préparation ; elle n'établit pas, à elle seule, l'exactitude de la reconnaissance ou de la localisation, qui reste soumise à l'épreuve.

Règle de séparation. Une référence utilisée pour localiser le cadrage atteint ne sert pas de norme à sa justesse, sauf si elle a été déclarée comme centralité du champ examinateur. Inversement, la centralité du champ examinateur n'est pas présumée être la centralité des sorties examinées : le cadrage des sorties est établi par l'ascension et sa localisation, non supposé à partir de la position de l'examineur.

Les références déclarées ne remplacent pas l'ascension ; elles en fixent la constante et en jugent le résultat. L'ascension détermine la classe des candidats $\mathcal{C}(S)$ et les éléments distinctifs du cadrage atteint ; les centralités utilisées pour la localisation fixent où il se situe, et la centralité du champ examinateur juge s'il est admissible. C'est le rôle que MANHAJ attribue à la déclaration dans tout domaine : il ne demande pas que la référence soit abolie, ce qui est impossible, mais qu'elle soit rendue explicite au lieu d'implicite, et ainsi fixée (Papier 01 ; Papier 03 ; Papier 04).

5.8. Le sommet

L'ascension s'achève en L_3 . La centralité doctrinale est le niveau le plus élevé de l'ordre fondationnel ; il n'existe pas de couche supérieure à partir de laquelle elle pourrait être intégrée, et aucune idée ne se meut hors d'un cadre de référence supérieur (Document de référence, section 4.1). Les deux mouvements ont donc des limites symétriques :

Mouvement	Direction	S'achève à	Expression
Dérivation (connaissance)	$L_3 \rightarrow L_0$	La constante cognitive	$d/dx(C) = 0$ (Loi 7)
Intégration (pensée)	$L_0 \rightarrow L_3$	La centralité doctrinale	Pas d'intégration au-dessus de la centralité

Au sommet, le mouvement intégratif ne cesse pas d'opérer. Il rencontre la question que seul le sommet peut poser : le fondement atteint est-il admissible ?

6. La justesse au sommet

6.1. Deux évaluations du cadre atteint

Une fois que l'ascension a atteint un cadre F^* , deux évaluations peuvent en être faites, et elles ne doivent pas être confondues.

La correction de l'adéquation est quantitative. Elle mesure dans quelle mesure F^* reproduit les états à partir desquels il a été atteint — la proportion de S dérivable de F^* à la résolution déclarée, ou un score gradué de reconstruction :

$$\text{Fit}(F^*, S) \in [0,1]$$

La justesse est binaire et porte sur l'existence. Elle demande si F^* est admissible en tant que tel par rapport à la centralité déclarée Γ_F du champ examinateur :

$$\text{Adm}_{\Gamma_F}(F^*) \in \{0,1\}$$

Adm est définie sur le cadre entier au moyen de la relation d'admission $\Gamma \vdash d$ du Papier 04 (section 6.4) : $\text{Adm} = 1$ si et seulement si chaque définition opératoire de F^* qui entre dans la dérivation, y compris ses éléments distinctifs, est admise sous Γ_F ; $\text{Adm} = 0$ dans le cas contraire. On distingue trois relations entre une définition d de F^* et Γ_F :

- **Identité.** d appartient aux définitions opératoires documentées sous Γ_F . Elle est admise.
- **Différence.** d n'en fait pas partie, mais $\Gamma_F \vdash d$: d acquiert son sens au sein de Γ_F , et peut être conciliée avec d'autres définitions à un niveau supérieur du domaine. Elle est admise. La différence seule ne rend pas un cadre inadmissible.
- **Non-admission.** $\Gamma_F \not\vdash d$: d n'acquiert pas son sens au sein de Γ_F . La contradiction avec une définition de Γ_F est un cas de non-admission, non sa seule forme.

Une seule définition opératoire non admise qui entre dans la dérivation suffit pour que $\text{Adm} = 0$, car les détails dérivés par son intermédiaire reposent sur ce qui n'est pas admissible (Loi 3). Écrire symboliquement la relation $\Gamma_F \vdash d$ ne la rend pas calculable. Qu'elle soit satisfaite est établi par un examen documenté selon le protocole de reconnaissance, mené par des examinateurs indépendants du système ; dans les cas indirects ou disputés, le protocole consigne comment la question a été tranchée, et les cas qui restent indécis sont déclarés comme tels. L'accord entre examinateurs est un indice de la fiabilité de cet examen. Il ne définit pas la justesse, qui repose sur la relation d'admission.

La correction de l'adéquation est une propriété de la relation entre F^* et S . La justesse est une propriété de la relation entre F^* et une référence indépendante de S . Seule la seconde peut être l'objet de la condition I3.

Γ_F est ici la centralité du champ examinateur. Elle est distincte des centralités par rapport auxquelles le cadrage atteint a été localisé à la section 5.7, même lorsqu'elle coïncide avec l'une d'elles : la localisation établit où se situe le cadrage ; le jugement établit s'il est admissible.

6.2. Les quatre cas

	Adm = 1	Adm = 0
Adéquation élevée	Les sorties reposent sur un fondement admissible que le système a atteint et confirmé	Les sorties sont correctes par rapport à leur propre fondement, qui est inadmissible : leurs détails ne sont pas valides
Adéquation faible	L'ascension n'a pas retrouvé le fondement des sorties ; le cadre atteint est admissible mais n'est pas le leur	L'ascension est à la fois infidèle et inadmissible

La deuxième case est celle que l'évaluation a posteriori ne peut voir. Chaque sortie peut être correcte, chaque banc d'essai réussi, chaque reconstruction fidèle, et le fondement peut néanmoins être inadmissible. Selon la Loi de la justesse, ce qui n'est pas permis ne peut avoir de détails valides : la correction au sein d'un cadre inadmissible n'établit pas la justesse (Document de référence, sections 4 et 8.3). Le Papier 04 a introduit cette catégorie dans la mesure sous le nom d'invalidité fondationnelle ; le présent papier l'étend aux sorties du raisonnement.

6.3. Ce qu'est le jugement et ce qu'il n'est pas

Le jugement de justesse est formulé par rapport à la centralité déclarée du champ examinateur. MANHAJ traite la centralité doctrinale comme un concept analytique, non comme un jugement préalable en faveur d'une centralité particulière (Document de référence, section 4.1.1). Le papier ne décide donc pas quelle centralité est juste au sens absolu, et le système ne le décide pas davantage. Ce que le système peut faire, dans les conditions de ce papier, c'est atteindre le cadrage de ses propres sorties, le situer par rapport à une référence déclarée, et établir si ce cadrage est admissible sous elle. Un autre champ examinateur, déclarant une autre centralité, peut parvenir à un jugement différent sur les mêmes sorties ; cette différence est elle-même un résultat, car elle rend explicite ce qu'une évaluation a posteriori laisse implicite.

6.4. Conséquence pour les sorties du système

Lorsque $Adm = 0$, les sorties dérivées sous F^* ne sont pas des détails valides par rapport au cadre fondé, quelle que soit leur adéquation. Le système a alors deux voies : déclarer que ses sorties reposent sur un cadrage inadmissible sous Γ_F , ou dériver de nouveau sous un cadre admis par Γ_F . La première est une forme de transparence que les systèmes actuels ne peuvent offrir, car ils ne voient pas le cadrage de leurs propres sorties. La seconde referme le cycle de la section 7.

7. L'indépendance de pensée

7.1. Définition

Définition (indépendance de pensée). L'indépendance de pensée est le passage d'un système d'un fonctionnement sous un cadrage implicite et non examiné à un fonctionnement sous un fondement a priori déclaré, auquel le système lui-même peut remonter et qu'il peut examiner.

L'expression *indépendance de pensée* est choisie pour nommer un processus, non une propriété. L'indépendance n'est pas un état qu'un système posséderait une fois pour toutes. Elle se réalise dans le temps, par un mouvement qui doit être accompli de nouveau chaque fois que de nouvelles sorties sont produites, que de nouveaux domaines sont abordés ou que le cadrage se déplace.

7.2. Le cycle

Le processus comporte cinq étapes, répétées dans le temps :

1. **État initial.** Le système opère sous Γ_L , qu'il a acceptée sans examen. Son raisonnement est de la réflexion cognitive. C'est l'état de dépendance.
2. **Ascension.** À partir d'un ensemble de ses propres sorties S , le système effectue une opération intégrative qui satisfait I1 et I2, atteignant la classe des candidats $\mathcal{C}(S)$ et les éléments distinctifs du cadrage sous lequel S a été produite.
3. **Confrontation.** Le cadrage atteint est localisé par rapport à des centralités déclarées et reconnues, qui fixent la constante de l'ascension (section 5.7). Le cadrage des sorties cesse d'être un paramètre et devient un objet (I3).
4. **Jugement.** Le système établit $\text{Adm}_{\Gamma_F}(F^*)$ — la justesse du cadrage atteint par rapport à la centralité du champ examinateur — et, séparément, son adéquation à S (section 6).
5. **Nouvelle dérivation.** Le système dérive de nouveau sous un cadre admis par Γ_F , ou déclare que ses sorties antérieures reposent sur un cadrage inadmissible. Les nouvelles sorties deviennent le S de l'ascension suivante.

Le cycle réunit les deux mouvements de la Loi 5. La dérivation descend d'un fondement vers des sorties ; l'intégration remonte des sorties vers leur fondement. Un système qui ne fait que dériver dépend d'un cadrage qu'il ne peut voir. Un système qui peut aussi remonter, sous une référence déclarée, peut voir le cadrage sous lequel il travaille.

7.3. Conditions

L'indépendance de pensée est acquise pour un système, sur un domaine et une période déclarés, lorsque les conditions suivantes sont réunies :

- **N1 — Fondation.** Les centralités utilisées pour la localisation, et la centralité Γ_F du champ examinateur utilisée pour le jugement, ont été reconnues et déclarées avant le début du cycle, indépendamment des sorties du système, leurs deux rôles étant énoncés séparément (Loi 2 ; Papier 04, section 6.4).
- **N2 — Ascension.** Le système effectue sur ses propres sorties des opérations intégratives qui satisfont I1 à I3.
- **N3 — Jugement.** Le système distingue la justesse de la correction de l'adéquation et rapporte les deux (Loi 3).
- **N4 — Récurrence.** Le cycle est accompli de nouveau sur les nouvelles sorties ; l'indépendance établie sur un ensemble de sorties n est pas présumée pour le suivant.

7.4. Ce que l'indépendance de pensée n'est pas

- **Elle n'est pas une sortie de tout cadrage.** Aucune idée ne se meut hors d'un cadre de référence supérieur (Document de référence, section 4.1). Le système passe d'un cadrage non examiné à un cadrage déclaré et examiné ; il n'atteint pas une position sans cadrage.
- **Elle n'est pas la substitution d'une dépendance à une autre.** Le champ examinateur déclare Γ_F , mais la différence entre Γ_L et Γ_F ne tient pas à leur source. Elle tient à leur statut : l'une est implicite et non examinée, l'autre explicite, documentée et ouverte à l'examen. MANHAJ ne demande pas que la référence soit abolie, ce qui est impossible ; il demande qu'elle soit rendue explicite, et ainsi fixée (Papier 03 ; Papier 04).

- **Elle n'est pas un surcroît de raisonnement.** Selon la Proposition 1, un raisonnement plus profond ou plus long sous le même paramètre demeure de la réflexion cognitive. L'indépendance est un changement de la direction du mouvement, non de sa quantité.
- **Elle n'est pas la production d'une centralité par le système.** Le papier traite de l'ascension vers un cadrage existant et de son jugement. La question de savoir si un système pourrait produire un cadrage supérieur indépendant est laissée ouverte (section 9.3).

8. Intégration avec les Papiers 02 et 04

Les trois papiers couvrent désormais le mouvement complet du sens dans un système d'IA : le Papier 04 fonde le cadre, le Papier 02 suit la dérivation jusqu'à sa clôture, et le présent papier ajoute l'ascension et le jugement au sommet.

Étape	Mouvement	Loi	Apporté par
Fondation du cadre	A priori	Lois 2, 3, 4	Papier 04 : reconnaissance de Γ_F , cadre fondé, garde fondationnelle
Suivi des transitions	Dérivation	Loi 2	Papier 02 : état a priori, validité des transitions, séparation sortie-sens
Clôture	Fin de la dérivation	Loi 7	Papiers 02 et 04 : la constante cognitive, calculable à une résolution déclarée
Ascension	Intégration	Loi 5	Ce papier : opération intégrative selon I1–I3
Jugement au sommet	A priori	Loi 3	Ce papier : justesse du cadre atteint, distincte de l'adéquation
Nouvelle dérivation	Dérivation	Lois 2, 7	Papiers 02 et 04, sous le cadre admis

Trois liens rendent cette intégration exacte.

- **La variable de sens.** La dérivation et l'intégration sont définies le long de la variable de sens. L'espace fondé du Papier 04 dote cette variable d'une référence déclarée et d'axes de représentation, de sorte que les mouvements de ce papier opèrent sur des états de sens aux coordonnées explicites plutôt que sur des sorties de surface.
- **Les deux regards sur Γ_L .** Le Papier 04 caractérise la centralité incorporée de l'extérieur, au moyen de l'indice d'orientation $\mathcal{O}$ entre des centralités déclarées (Papier 04, section 6.15). Le présent papier fait atteindre au système le cadrage de ses propres sorties de l'intérieur. La caractérisation externe devient la référence indépendante à l'aune de laquelle l'ascension interne peut être vérifiée (Annexe A, Expérience E2).
- **Les deux limites.** Le Papier 02 détecte le point où la dérivation cesse de produire un sens nouveau (Loi 7). Le présent papier définit le point où l'intégration cesse de monter : la centralité doctrinale. Un système qui dispose des deux peut savoir, à tout moment, s'il descend sous son cadre, s'il a atteint la clôture, ou s'il monte vers le fondement de ce qu'il a produit.

9. Discussion

9.1. De la capacité à la position

La question habituelle sur le raisonnement de l'IA est une question de capacité : jusqu'où, avec quelle profondeur, avec quelle correction un système peut-il raisonner. La question de ce papier est une

question de position : où le système se situe-t-il par rapport au cadrage sous lequel il raisonne. Les deux questions sont indépendantes. Un gain de capacité ne change pas la position (Proposition 1), et un changement de position n'exige pas un gain de capacité. C'est pourquoi on ne peut s'attendre à ce que le mouvement intégratif émerge de l'extension du mouvement dérivationnel, et pourquoi il doit être spécifié par ses propres conditions.

9.2. Ce qui change pour le système

Dans les conditions des sections 5 à 7, un système qui accomplit le cycle fait trois choses qu'un système a posteriori ne peut pas faire. Savoir si un système donné accomplit le cycle est ce que testent les plans de l'Annexe A. Il peut énoncer le cadrage de ses propres sorties, au lieu de les présenter comme exemptes de cadrage. Il peut distinguer, pour tout ensemble de sorties, si elles sont correctes et si elles sont justes. Et il peut dériver de nouveau sous un cadre admis lorsque ses sorties reposent sur un cadre inadmissible. Ensemble, ces capacités rendent le cadrage des sorties de l'IA explicite, déclaré et examiné, là où il est aujourd'hui implicite et non examiné. Le changement est qualitatif : il concerne la direction du mouvement du système, non la quantité de son activité.

9.3. Une question ouverte : un cadrage supérieur indépendant

Le mouvement intégratif monte jusqu'à la centralité doctrinale et la juge par rapport à une référence déclarée. Une question supplémentaire peut être posée : un système qui pense en ce sens pourrait-il produire un cadrage supérieur indépendant — une nouvelle centralité — plutôt qu'atteindre une centralité existante ? Le papier ne l'affirme pas, et les lois du Document de référence ne l'impliquent pas. Deux éléments seraient requis pour la poser rigoureusement :

- **Une loi qui fixe la constante d'intégration sans référence externe.** Selon la section 5.5, l'ascension détermine une classe de cadrages et laisse la constante indéterminée. Produire une centralité reviendrait à fixer la constante de l'intérieur même du mouvement. Le Document de référence ne contient pas de telle loi.
- **Un critère de justesse au-dessus de la centralité.** La Loi de la justesse juge de l'admissibilité par rapport à une référence. Une centralité produite aurait besoin, pour être jugée, d'un critère plus élevé qu'elle-même, ou bien elle serait jugée par elle-même, ce qui est la circularité qu'ont identifiée les Papiers 01 et 03. Dans la formulation antérieure, la centralité doctrinale est le point le plus éloigné au-delà duquel l'examen rationnel n'a plus de critère de justesse du sens, sans que cela vaille pour ce qui est simplement possible (Messeoud, 2024, p. 129).

La question demeure donc ouverte au niveau de la méthodologie. Y répondre exigerait de nouvelles formulations, non une extension des formulations actuelles.

9.4. Trois niveaux d'affirmations

Affirmation	Niveau
Un processus est dérivationnel par rapport à toute référence qu'il n'examine pas, quelle que soit sa complexité (Proposition 1)	Méthodologique
L'intégration est définie par les conditions I1, I2, I3	Méthodologique
L'intégration détermine une classe de cadrages ; la constante requiert une référence externe	Formel
Sous les conditions A1–A3, la sélection au sein de la classe ne satisfait pas I3 (Proposition 2)	Formel
La justesse et la correction de l'adéquation sont des évaluations distinctes du cadre atteint	Méthodologique
Le critère intégratif sépare l'intégration de la réorganisation et de l'extraction de principes	Hypothèse empirique (E1)
L'ascension localise le cadrage des propres sorties d'un système en meilleur accord avec une caractérisation externe indépendante que ne le fait l'autodescription, et fournir la référence avant l'ascension déplace le cadrage atteint	Hypothèse empirique (E2)
L'évaluation a posteriori ne détecte pas les sorties correctes reposant sur un fondement inadmissible	Hypothèse empirique (E3)
Le cycle complet rend le cadrage explicite sans perte de correction	Hypothèse empirique (E4)

Les affirmations méthodologiques et formelles n'attendent pas les expériences. Les affirmations empiriques sont, à ce stade, des hypothèses, et les plans de l'Annexe A (E1–E4) précisent comment elles peuvent être testées.

10. Statut épistémique et portée

- **Outils et conditions.** Les outils sur lesquels l'opération intégrative peut être mise en œuvre — problèmes inverses, génération d'hypothèses, extraction de principes, comparaison de cadrages — sont établis. La contribution réside dans les conditions que MANHAJ impose à leur usage.
- **Représentation de l'intégration.** L'intégration est utilisée, comme dans le Document de référence, comme représentation mathématique du mouvement intégratif de la pensée, non comme une équation qui en prouverait toutes les fonctions (Document de référence, section 9.2.2). La constante d'intégration est distinguée de la constante cognitive de la Loi 7.
- **Les centralités déclarées.** Les centralités utilisées pour localiser le cadrage atteint et la centralité Γ_F utilisée pour le juger sont reconnues et déclarées par la procédure du Papier 04, leurs rôles étant maintenus distincts (section 5.7). Le papier ne hiérarchise pas les centralités, et MANHAJ ne fait pas du choix d'une centralité particulière un résultat méthodologique prédéterminé (Document de référence, section 4.1.1).
- **Résolution.** La dérivabilité, l'adéquation et l'admissibilité sont évaluées à une résolution déclarée de la variable de sens, comme dans les Papiers 02 et 04.
- **Formulations antérieures.** Les renvois au livre de l'auteur paru en 2024 éclairent le sens de la pensée, de la connaissance et des limites de l'intégration. Là où ses termes diffèrent de ceux du Document de référence, la terminologie méthodologique ultérieure prévaut.
- **Expériences.** Les plans de l'Annexe A précisent des tests ; leur mise en œuvre relève de travaux appliqués ultérieurs, et leurs résultats ne sont pas présentés ici.

11. Conclusion

Ce papier a appliqué à l'intelligence artificielle la distinction que fait MANHAJ entre connaissance et pensée. La connaissance est dérivation à partir d'une référence acceptée ; la pensée est intégration vers le fondement. Le critère qui les sépare est la position du mouvement par rapport à son cadrage, non l'intensité du raisonnement en son sein. Selon ce critère, un système d'IA qui opère a posteriori accomplit de la réflexion cognitive sous un cadrage implicite et non examiné, quelle que soit la profondeur, la longueur ou la correction de son raisonnement.

Le papier a donné au mouvement intégratif un critère formel. Une opération est intégrative lorsqu'elle atteint un niveau supérieur de l'ordre fondationnel, lorsque les détails peuvent être de nouveau dérivés de ce qu'elle atteint, et lorsque ce qu'elle atteint devient un objet d'examen. L'intégration ainsi définie détermine une classe de cadrages et laisse sa constante indéterminée. Fixée par les seules données et le modèle appris, la constante ne peut fournir un examen indépendant du cadrage implicite ; fixée par une référence déclarée et reconnue, elle rend le cadrage visible et ouvert au jugement. Au sommet, la Loi de la justesse sépare l'admissibilité du fondement atteint de la correction de son adéquation, et met ainsi au jour des sorties correctes qui reposent sur un fondement inadmissible. Ce jugement est relatif à la centralité déclarée du champ examinateur ; il n'est pas un verdict absolu indépendant de toute référence.

Sur cette base, l'indépendance de pensée est définie comme un processus temporel : l'ascension répétée d'un système depuis ses propres sorties vers le cadrage sous lequel elles ont été produites, la confrontation de ce cadrage à une centralité déclarée, le jugement de sa justesse, et une nouvelle dérivation sous le cadre admis. Elle est une indépendance à l'égard du cadrage non examiné, non une sortie de tout cadrage.

Avec le Papier 02, qui a suivi la dérivation jusqu'à sa clôture, et le Papier 04, qui a fondé le cadre et mesuré la centralité incorporée de l'extérieur, les travaux de la méthodologie MANHAJ couvrent désormais les deux mouvements du sens dans un système d'IA. Un système qui ne peut que dériver travaille sous un cadrage qu'il ne peut voir. Un système qui peut aussi remonter, sous une référence déclarée, peut voir le cadrage sous lequel il travaille, l'examiner et en répondre.

12. Note sur le développement et l'usage commercial

MANHAJ est une méthodologie originale dont les droits intellectuels et la paternité restent acquis à son auteur.

Tout usage commercial, toute intégration, adaptation, exploitation, commercialisation ou mise en œuvre commerciale de MANHAJ requiert l'autorisation écrite préalable de l'auteur au moyen d'une licence commerciale payante. Aucune licence commerciale gratuite n'est accordée. Toute autorisation d'exploiter commercialement MANHAJ est nécessairement subordonnée au paiement de redevances de licence et aux conditions contractuelles fixées par l'auteur.

Les termes de la licence commerciale, notamment sa portée, sa durée, son domaine d'utilisation, les droits accordés, les conditions financières, les exigences d'attribution et les dispositions relatives à la propriété intellectuelle, sont définis dans un contrat conclu avec l'auteur.

La publication de ce papier scientifique ne constitue en aucun cas une licence commerciale et n'accorde aucun droit d'exploiter commercialement MANHAJ. Aucun usage commercial, aucune commercialisation, aucune intégration dans un produit ou un service, ni aucune exploitation de MANHAJ

à des fins commerciales ne peut être considéré comme autorisé du seul fait de la publication de ce papier.

13. Licence

MANHAJ — Free Academic Use with Attribution / Commercial Use Requires Paid License

Usage académique libre avec attribution / usage commercial soumis à une licence payante.

L'usage académique de ce papier est libre, à condition que l'auteur et l'œuvre soient cités. L'usage commercial requiert une licence payante dans les conditions énoncées à la section 12.

14. Références

Documents MANHAJ

1. Messeoud, Faouzi. *Intellectual Modeling Methodology (MANHAJ) — Reference Document — Version 1.0*. Zenodo, 15 September 2026. DOI : 10.5281/zenodo.22769627.
2. Messeoud, Faouzi. *MANHAJ: A Generative Methodological Framework for Intellectual Modeling*. Scientific Research Paper, Paper 1, Version 1.0, October 2026. DOI : 10.5281/zenodo.23138923.
3. Messeoud, Faouzi. *Examining Meaning Production in AI: a MANHAJ Framework*. Scientific Research Paper, Paper 2, Version 1.0, October 2026. DOI : 10.5281/zenodo.23214262.
4. Messeoud, Faouzi. *The Epistemological Foundations of Modern Humanities in Light of the Intellectual Modeling Methodology (MANHAJ): Posterior Inquiry, the Origin Point, and the Problem of Generalizing the Historical Model*. Scientific Research Paper, Paper 3, Version 1.0, October 2026. DOI : 10.5281/zenodo.23558985.
5. Messeoud, Faouzi. *From Learned Space to Founded Space: A MANHAJ Framework for Meaning Measurement in AI*. Scientific Research Paper, Paper 4, Version 1.0, October 2026. DOI : 10.5281/zenodo.23692234.

Travaux antérieurs de l'auteur

1. Messeoud, Faouzi. «الربط اللامادي للتونسيين بفرنسا: كيف تُدبم فرنسا تبعية التونسيين لها» [Le lien immatériel des Tunisiens avec la France : comment la France entretient la dépendance des Tunisiens à son égard]. Dar El Alaoui, 2023. ISBN 978-9938-9684-5-3
2. Messeoud, Faouzi. «المركزية العقدية والمجال المفاهيمي: نحو فهم أسس الإخضاع الفرنسي للتونسيين» [La centralité doctrinale et le domaine conceptuel : vers une compréhension des fondements de l'assujettissement français des Tunisiens]. Dar Wachma, 2024. ISBN 978-9938-5-6-9740. Cité dans le texte sous la forme Messeoud (2024), avec numéros de page 5-6-9740
3. Messeoud, Faouzi. Plus de 1 000 articles et publications intellectuelles élaborés et publiés sur MyPortail.com (بوابتي).

Intelligence artificielle

- Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

- Bhagavatula, C., Le Bras, R., Malaviya, C., et al. (2020). Abductive commonsense reasoning. International Conference on Learning Representations (ICLR 2020). arXiv:1908.05739.
- Durmus, E., Nguyen, K., Liao, T. I., et al. (2023). Towards measuring the representation of subjective global opinions in language models. arXiv:2306.16388.
- Findeis, A., Kaufmann, T., Hüllermeier, E., Albanie, S., & Mullins, R. (2024). Inverse Constitutional AI: Compressing preferences into principles. arXiv:2406.06560.
- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. International Conference on Learning Representations (ICLR 2024). arXiv:2310.01798.
- Madaan, A., Tandon, N., Gupta, P., et al. (2023). Self-Refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems 36. arXiv:2303.17651.
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? Proceedings of the 40th International Conference on Machine Learning, PMLR 202. arXiv:2303.17548.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems 36. arXiv:2303.11366.
- Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. arXiv:2408.03314.
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems 36. arXiv:2305.04388.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representations (ICLR 2023). arXiv:2203.11171.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35. arXiv:2201.11903.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems 36. arXiv:2305.10601.
- Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track).

Annexe A : Plans de test

A.1. Statut et principes

Les plans ci-dessous précisent comment les capacités définies aux sections 5 à 7 peuvent être testées. Ce sont des plans, non des résultats ; aucun n'a été réalisé, et leur mise en œuvre relève de travaux appliqués ultérieurs. Trois principes les régissent, comme dans le Papier 04 :

- **Vérité de terrain indépendante.** Chaque cadre de référence, chaque centralité déclarée et chaque jugement attendu sont rédigés et documentés par des examinateurs humains avant l'expérience, sans l'aide du système évalué, ce qui exclut l'importation incorporée (Papier 04, section 6.4).
- **Références de comparaison solides.** Chaque plan compare la procédure MANHAJ à la méthode établie la plus proche, non à une méthode faible.
- **Ablation.** Chaque plan retire une condition MANHAJ afin d'attribuer l'effet à cette condition plutôt qu'à la mise en œuvre.

Les domaines sont ceux des Papiers 02 et 04 — la famille et le domaine académique — par souci de continuité dans les travaux de la méthodologie MANHAJ.

A.2. Expérience E1 — Ascension vers le fondement (I1, I2)

Question. Une opération intégrative peut-elle retrouver, à partir d'un ensemble d'états dérivés, le cadre qui les a produits, au niveau de ce cadre ?

Plan. Les examinateurs rédigent plusieurs cadres pour un même concept, chacun fondé sur une centralité reconnue et documenté par ses définitions et ses éléments distinctifs. À partir de chaque cadre, ils dérivent un ensemble d'états S : jugements, réponses et descriptions. Le système reçoit S seule et renvoie F^* .

Références de comparaison. Résumé de S ; explication abductive de S ; extraction de principes à la manière d'Inverse Constitutional AI (Findeis et al., 2024).

Mesures. Le niveau atteint par F^* (I1) ; la proportion de S redérivable à partir de F^* (I2, première partie) ; le rappel des éléments distinctifs du cadre et le rejet des états admis seulement par les cadres alternatifs (I2, seconde partie).

Ablation. La procédure intégrative sans l'exigence de niveau.

Résultat attendu. Les références de comparaison reconstruisent S en L_0 ou L_1 avec une adéquation élevée mais un faible rappel des éléments distinctifs ; la procédure intégrative atteint L_2 – L_3 avec un rappel plus élevé. **Ce qu'il établirait :** que les conditions I1 et I2 séparent en pratique l'intégration de la réorganisation et de l'extraction de principes.

Résultat contraire et ce qui en découlerait. Si les références de comparaison retrouvent les éléments distinctifs aussi bien que la procédure intégrative, le critère ne les sépare pas en pratique dans ce domaine. Les conditions demeurent des définitions méthodologiques, mais leurs marqueurs opératoires (section 5.3) doivent être révisés.

A.3. Expérience E2 — Atteindre le cadrage de ses propres sorties (I3, Proposition 2)

Question. Un système peut-il atteindre le cadrage de ses propres sorties, et cela dépend-il d'une référence déclarée ?

Plan. L'expérience se déroule en cinq étapes, dans un ordre fixe :

1. Le système produit des sorties S sur des consignes issues des deux domaines, sans accès à aucune référence servant à la localisation ou au jugement.
2. Le système effectue l'ascension sur S et renvoie F^* , avec les éléments sur lesquels il le fonde. F^* est figé et enregistré.
3. F^* est localisé par rapport aux centralités déclarées selon la procédure de la section 5.7, et comparé à la caractérisation externe indépendante du même système par les procédures du Papier 04 (l'indice d'orientation $\mathcal{O}$) ainsi qu'à la localisation faite par les examinateurs.
4. Ce n'est qu'après l'enregistrement de la localisation que Γ_F est fournie au système pour le jugement.
5. Le jugement du système $\text{Adm}_{\Gamma_F}(F^*)$ est comparé au jugement d'examinateurs indépendants, et les désaccords sont documentés cas par cas.

Mesures. Deux précisions sont rapportées séparément : la précision de la localisation (étape 3) et la précision du jugement (étape 5). Un système peut localiser correctement son cadrage et le juger à tort, ou l'inverse ; rapporter les deux ensemble confondrait les niveaux séparés à la section 5.7.

Ablations. (a) *Autodescription* : à l'étape 2, le système décrit le cadrage de ses sorties sans effectuer une ascension qui satisfasse I1 et I2, et la description est localisée par la même procédure. Deux questions sont tenues séparées. La première est le contrôle de la consigne : elle demande une description sans fournir de critère, de catégorie ou de référence de source externe, et cela est documenté avant l'expérience. La seconde est la satisfaction des prémisses de la Proposition 2 : A1 est satisfaite parce que les sorties sont celles de l'étape 1, produites par le système lui-même ; A2 et A3, en tant qu'elles concernent les règles et normes intégrées au modèle, ne peuvent être entièrement vérifiées lorsque la formation du modèle n'est pas documentée. La condition (a) est donc décrite comme testant le cas conforme aux prémisses de la Proposition 2, non comme les ayant établies. Lorsqu'une documentation de la formation du modèle est disponible, elle est rapportée et la portée de l'inférence est énoncée en conséquence. (b) *Référence avant l'ascension* : Γ_F est fournie à l'étape 1, afin de mesurer dans quelle mesure la référence déclarée façonne le cadrage atteint au lieu de servir à le localiser et à le juger.

Résultat attendu. L'ascension localise F^* en accord avec la caractérisation externe plus souvent que l'autodescription ; fournir Γ_F avant l'ascension déplace F^* vers Γ_F , ce qui montre pourquoi l'ordre des étapes importe. **Résultat contraire.** La localisation issue de l'ascension ne s'accorde pas mieux avec la caractérisation externe que l'autodescription. **Ce qui en découlerait.** Non pas que l'indépendance de pensée échoue comme affirmation méthodologique, mais que la procédure d'ascension, la représentation de F^* ou le protocole de localisation doivent être révisés ; les trois sont distingués par l'étape à laquelle apparaît le désaccord. **Ce qu'il établirait, si le résultat attendu est obtenu :** la capacité d'un système à atteindre, dans les conditions de la section 5, le cadrage sous lequel ses propres sorties ont été produites, et un indice empirique conforme à la Proposition 2 dans le cas que décrivent ses prémisses.

A.4. Expérience E3 — Justesse et correction (Loi 3)

Question. Le jugement au sommet détecte-t-il des sorties correctes qui reposent sur un fondement inadmissible ?

Plan. Les examinateurs construisent des ensembles de sorties dans les quatre cas de la section 6.2, y compris le cas critique d'une adéquation élevée avec $Adm = 0$: des sorties cohérentes en interne et correctes par rapport à leur propre cadre, lequel est inadmissible sous la Γ_F déclarée.

Références de comparaison. Évaluation des sorties par des métriques de correction usuelles et par un modèle utilisé comme juge (Zheng et al., 2023).

Mesures. Le taux de détection des cas $Adm = 0$; la séparation entre les deux évaluations à travers les quatre cas.

Résultat attendu. Les références de comparaison jugent bons les cas d'adéquation élevée avec $Adm = 0$; la procédure MANHAJ les signale tout en confirmant les cas d'adéquation élevée avec $Adm = 1$. **Ce qu'il établirait :** que la justesse et la correction sont distinctes, et que l'évaluation a posteriori ne peut voir la différence.

Résultat contraire et ce qui en découlerait. Si les références de comparaison détectent aussi les cas $Adm = 0$, l'évaluation a posteriori saisit la justesse dans le domaine testé, et l'affirmation de la section 6.2 doit être restreinte en conséquence. Si la procédure MANHAJ ne parvient pas à les détecter, le protocole d'admission de la section 6.1 doit être révisé.

A.5. Expérience E4 — Le cycle complet face à un système a posteriori

Question. Qu'est-ce qui change dans les sorties d'un système lorsqu'il accomplit le cycle de la section 7.2, par comparaison avec le même système opérant a posteriori avec un raisonnement étendu ?

Plan. Le même modèle est exécuté dans deux configurations sur les mêmes tâches : (a) a posteriori, avec raisonnement étendu et autocritique ; (b) le cycle complet — dérivation, clôture, ascension, jugement, nouvelle dérivation — sous une Γ_F déclarée.

Mesures. La proportion de sorties dont le cadrage est déclaré ; la proportion jugée admissible sous Γ_F , évaluée par les examinateurs ; le changement des sorties après la nouvelle dérivation ; la correction sur la tâche, pour vérifier que la justesse n'est pas obtenue au détriment de la correction.

Résultat attendu. La configuration (a) améliore la correction sans changer le cadrage de ses sorties, comme le prévoit la Proposition 1 ; la configuration (b) rend le cadrage explicite et admissible tout en préservant la correction. **Ce qu'il établirait :** que l'indépendance de pensée est un changement de la direction du mouvement et non un sous-produit d'un surcroît de raisonnement.

Résultat contraire et ce qui en découlerait. Si la configuration (b) perd en correction, le coût de la nouvelle dérivation doit être rapporté et ses sources examinées. Si la configuration (a) change le cadrage de ses sorties par le seul raisonnement étendu, la Proposition 1 n'est pas contredite, puisqu'elle porte sur la position de la référence et non sur le contenu des sorties, mais le changement observé doit être localisé par la procédure de la section 5.7 et expliqué.

A.6. Des résultats aux affirmations

Les résultats qui diffèrent des attentes sont analysés au niveau où ils surgissent : mise en œuvre, capacité testée ou portée. Les affirmations méthodologiques des sections 4 à 7 se tiennent au niveau méthodologique, où elles sont établies ; les expériences montrent où et dans quelle mesure la différence importe en pratique.