

Une mémoire persistante n'est pas une mémoire biographique

Un cadre de décomposition, deux constructions candidates, et un programme expérimental réfutable

Cinquième volet de la série Encodage, transduction et modèles du monde. La littérature sur la mémoire des agents traite comme une seule question ce qui en est plusieurs. Cet article sépare ce qu'elle confond, propose deux constructions opératoires, et spécifie les expériences qui les détruiraient.

1. Introduction

Le volet précédent, *Une représentation est une politique de conservation de l'information*, a caractérisé les architectures prédictives latentes comme nulles sur les axes de persistance mnésique et de modèle de soi. Ce volet traite ces axes.

Le statut de ce texte doit être posé d'emblée. ***Ce n'est pas une théorie stabilisée de la mémoire artificielle mais un cadre de décomposition assorti d'un programme expérimental.*** La partie robuste est négative : elle établit ce qui ne s'implique pas. Les deux constructions positives sont des candidates opératoires, dont les seuils sont conventionnels et déclarés comme tels. Un troisième niveau, souvent invoqué sous le nom d'autobiographie, est traité en §9 comme problème ouvert plutôt que comme résultat.

La contribution tient en deux distinctions, qui ne sont pas de même nature et qu'il serait malhonnête de présenter symétriquement.

La première est architecturale :

stockage ≠ récupération

Une information peut être écrite sans être restituable, et restituable sans être écrite, la reconstruction depuis des traces distribuées produisant le second cas.

La seconde est fonctionnelle, et c'est la contribution théorique :

récupération ≠ adaptation persistante ≠ attribution causale

Un magasin vectoriel restitue sans rien adapter. Un affinage adapte sans rien restituer. Un graphe d'audit enregistre une participation au calcul sans établir une contribution causale.

Une dissymétrie d'instrumentation en découle, et elle constitue le diagnostic principal. ***Le domaine instrumente abondamment le stockage et la récupération, capacité, latence, taux de rappel ; beaucoup moins l'adaptation persistante ; presque jamais l'attribution causale.*** L'effort de mesure décroît exactement dans l'ordre où la difficulté conceptuelle croît.

De ces distinctions découlent deux constructions candidates :

$M_E = \text{Ép} \wedge \text{Acc}$: mémoire épisodique, indépendamment de la source

$B_C(\tau, C) = M_E \wedge [\text{Source}(E) = \text{propre}] \wedge D_X \wedge [I(E; C) \geq \tau]$: mémoire biographiquement constitutive, au seuil τ sous le modèle causal C

Distinctes ne signifie pas indépendantes. L'attribution causale n'a de sens que si l'adaptation est établie, ce qui rend la structure conditionnelle plutôt que plate, et les deux paramètres explicites rendent la seconde construction relative plutôt que catégorielle.

Une discipline sur la portée s'impose. Quatre propositions circulent sans être distinguées : aucun système ne possède X ; aucun système publié ne possède X ; aucune publication n'a démontré X ; aucune publication n'a testé X sous un protocole approprié. Cet article établit la quatrième, sur un périmètre bibliographique explicitement illustratif. La revendication est donc étroite : **dans la littérature considérée ici, nous n'avons identifié aucune évaluation explicite de la conjonction de ces termes.**

Rien ne sera dit de l'autonoéticité ni de la qualité subjective du souvenir.

2. Deux non-implications symétriques

Première non-implication. Une définition tentante consiste à évacuer la phénoménologie au profit d'un critère contrefactuel : un épisode appartient à l'agent si sa suppression modifie durablement sa politique. Ce critère est falsifiable et trop permissif. Un réseau affiné une fois sur un épisode voit ses poids modifiés ; retirer l'épisode et réapprendre change le comportement, durablement, sans aucune récupération.

influence causale persistante d'un événement $\nRightarrow$ mémoire de cet événement

Une intoxication alimentaire modifie durablement mes préférences, une lésion modifie ma politique. On peut être profondément transformé par un événement dont on n'a aucune mémoire. L'effet mesuré peut par ailleurs transiter par une chaîne indirecte, $E \rightarrow D \rightarrow U \rightarrow \theta \rightarrow \pi$, où D est une donnée dérivée, U une mise à jour, θ les paramètres et π la politique : une ablation détecte un effet distal, elle ne démontre pas qu'il existe une trace de E .

Seconde non-implication. Corriger cela en exigeant que tout souvenir ait transformé l'agent produit une définition trop étroite.

mémoire d'un événement $\nRightarrow$ transformation par cet événement

Je me souviens d'un déjeuner insignifiant d'il y a cinq ans. Je le discrimine, je le restitue, je sais qu'il m'est arrivé. Il n'a modifié aucune croyance, aucune préférence, aucune politique détectable. Une définition qui refuse d'appeler cela un souvenir a défini autre chose que la mémoire.

Les deux ensemble commandent la structure : la transformation ne sépare pas la mémoire de la non-mémoire, elle sépare M_E de B_C .

3. Exposition, traitement, intégration

Un préalable de typage évacue une confusion tenace. Les taxonomies additives de la forme $M = M_{\text{contexte}} + M_{\text{externe}} + M_{\text{poids}} + M_{\text{état}}$ mélangent des **supports**, qui désignent où l'information réside, et des **contenus**, qui désignent ce qu'elle représente. Une procédure peut résider dans les poids, dans un magasin externe ou dans un outil. La conséquence utile est qu'aucun substrat élémentaire ne suffit et que la propriété recherchée appartient au système couplé.

Un second préalable concerne le trajet d'une entrée :

1. **Exposition** : E appartient à l'histoire causale de l'agent. Un paquet réseau, un message indésirable, une corruption mémoire satisfont ce régime.
2. **Traitement** : E traverse le pipeline et influence une activation quelconque. Presque toute entrée satisfait ce régime, qui est donc sans valeur discriminante.
3. **Intégration fonctionnelle** : E est accessible à au moins une composante décisionnelle ou mnésique au-delà du prétraitement, indépendamment de son écriture éventuelle.

Le stockage est une quatrième question, orthogonale : une entrée intégrée peut n'être jamais écrite, une entrée écrite peut n'avoir jamais été intégrée.

Cette séparation empêche la circularité qui menace toute définition de l'expérience propre. La définir par ce que le système écrit en mémoire produit une tautologie : serait expérience propre ce que la politique mnésique décide de traiter comme tel. En l'ancrant dans l'intégration fonctionnelle, la définition devient indépendante de la politique de mémoire.

4. Les termes

Chaque terme désigne une propriété du système. Chaque terme possède un estimand distinct, qui est ce que l'expérimentateur peut mesurer et qui n'est jamais la propriété elle-même. Confondre les deux ferait dépendre ce qu'un système **est** de ce que l'expérimentateur a réussi à mesurer.

Terme	Propriété	Estimand	Limitation
Ép	indexation comme occurrence particulière	test d'adressage de l'occurrence	dépend de l'instrumentation
Acc	discriminabilité, récupérabilité	fidélité mnésique contre trace encodée	trace de référence partielle
Source	origine de l'événement et de sa mémoire	traversé contre rapporté ; robustesse à l'induction	accès aux journaux d'intégration
D_X	déplacement persistant de disposition	effet détecté sur un domaine de sondage	domaine ouvert, non couvrable
I	contribution causale de E à ΔS	classe d'estimands, relative à un modèle causal	indéfini si D_X est nul

Ép, épisodicité

L'épisodicité est une propriété représentationnelle, distincte des propriétés fonctionnelles qui suivent, et les fondre dans un terme unique masquerait qu'elles se dissocient. Un système peut être parfaitement accessible sans être épisodique : il restitue le numéro atomique du sodium sans qu'aucune occurrence soit en jeu.

Ép(E) = 1 lorsque le contenu est indexé comme occurrence particulière dans l'histoire d'observation de l'agent, adressable en tant que telle indépendamment de tout contenu général qu'elle aurait pu véhiculer.

Le cas de l'acquisition de connaissance en fournit le test le plus lisible : le système distingue-t-il « P est vrai » de « j'ai appris P », l'événement d'acquisition étant adressable indépendamment de la proposition acquise ? Un enregistrement associant une proposition à un pointeur de source ne passe pas ce test, puisqu'il attache une provenance à un fait sémantique.

Ce test est exemplaire, non général. Une action motrice, une interaction sociale, une séquence prolongée, un événement multimodal sans proposition acquise relèvent tous de l'épisodicité sans qu'aucun contenu propositionnel soit disponible pour la comparaison. Le critère général reste l'adressabilité de l'occurrence ; sa vérification demande, pour ces classes, des tests spécifiques qui n'existent pas encore et dont la construction fait partie du programme.

Acc, accessibilité

Acc(E) = (discriminabilité, récupérabilité)

La discriminabilité est la capacité à ne pas confondre E avec un épisode voisin ; elle n'exige pas un identifiant stable. La récupérabilité est la restitution du contenu sur sollicitation, par lecture ou par reconstruction, et n'exige pas la persistance d'un index, la mémoire épisodique biologique étant largement reconstructive [18]. L'ordonnancement relatif n'est pas requis ; le caractère passé de l'épisode est acquis par construction. La rapportabilité est écartée : l'exiger réintroduirait des capacités linguistiques que ce cadre entend éviter.

L'ouverture à la reconstruction crée un risque qu'il faut fermer, un modèle génératif produisant une version plausible mais fausse passant sinon le test. La fidélité porte donc sur trois objets, et non deux :

E_monde → E_encodé → Ê_restitué

F_perception = sim(E_monde, E_encodé) et F_mémoire = sim(E_encodé, Ê_restitué)

Seule la seconde est une fidélité mnésique. Un agent ayant mal perçu un événement puis restituant fidèlement sa perception erronée possède une excellente mémoire d'une mauvaise observation. La grandeur n'est par ailleurs pas scalaire : une reconstruction peut être lexicalement éloignée et causalement exacte, ou sémantiquement proche en attribuant l'action au mauvais agent. D'où $F = (F_{entité}, F_{relation}, F_{temporel}, F_{action}, F_{source})$, avec pondération déclarée par domaine ; en contexte clinique, F_{action} et F_{source} dominent.

Une limite d'applicabilité doit être assumée. *E_encode* n'est accessible que dans les architectures qui journalisent leur encodage. Sur un système opaque, la seule mesure disponible est la distance à l'événement, qui mélange perception et mémoire. ***Le protocole n'est donc pas architecturalement agnostique : il fonctionne sur des systèmes instrumentés et dégrade sur les autres.***

Source

La source est une caractérisation orthogonale, non une condition de la mémoire épisodique. Un agent peut avoir une mémoire épisodique parfaitement constituée d'un événement qu'il a observé sans en être acteur, et une définition qui l'interdit confond mémoire et participation.

Deux provenances doivent être distinguées, ce qu'un terme unique écrasait dans les versions antérieures de ce cadre :

provenance de l'événement : l'agent était-il acteur, observateur, ou destinataire d'un récit ?

provenance de la mémoire : la trace résulte-t-elle de l'observation de l'agent, ou d'une écriture externe ?

Un agent qui lit un document possède une mémoire propre d'un événement externe : la provenance de la mémoire est propre, celle de l'événement ne l'est pas. Lire son propre journal produit un épisode de lecture, distinct des épisodes journalisés. Recevoir une transcription de ses interactions crée un épisode de réception, non les épisodes transcrits.

Source(E) ∈ {propre, externe observée, importée}

Seule la première catégorie entre dans *B_C*, une biographie portant sur ce que l'agent a traversé. Le critère est l'intégration fonctionnelle au sens de §3, et sa minimalité doit être assumée : une donnée de contexte injectée par un opérateur et utilisée une fois pour choisir une action relève de la trajectoire d'interaction de l'agent. ***Ce terme ne signifie pas « vécu » au sens autobiographique, mais « causalement incorporé dans la boucle de l'agent ».***

D_X, déplacement persistant de disposition

L'épisode modifie la disposition de l'agent, et cette modification survit à la réinitialisation du contexte. Le terme n'est pas « constitution d'état », qui promettrait un accès interne que la définition refuse. Ce qui est mesuré est un déplacement de disposition, défini directement sur les politiques et sans référence à un état latent :

$$**\pi_i \sim_X \pi_j \Leftrightarrow \forall x \in X, \pi_i(x) \approx \pi_j(x)**$$

Le quantificateur universel porte sur la propriété. L'estimand porte sur un domaine de sondage fini, ce qui interdit une phrase courante : on ne peut pas conclure qu'un épisode n'a pas modifié l'agent, seulement qu'***aucune modification n'est détectée sur X_test***. La disposition possède un domaine de validité, exactement comme l'invariance dans le volet précédent.

I, contribution causale, conditionnelle à D_X

Deux choses que les graphes de provenance confondent systématiquement doivent être séparées, et cette distinction est la plus importante de l'article :

P_trace : E a participé au calcul qui a produit ΔS

I : E a contribué causalement à ΔS

Un moteur qui enregistre « j'ai utilisé E_7 lors de cette mise à jour » produit une provenance procédurale. L'utilisation n'implique pas la contribution : E_7 peut avoir été traité sans rien déterminer. Se satisfaire de P_trace reviendrait à commettre, au niveau de l'audit, l'erreur que §2 dénonce au niveau de l'ablation.

I ne peut pas être défini par la nécessité individuelle de l'épisode, sous peine de perdre les épisodes redondants : si E_1 et E_2 conduisent chacun à la croyance X , retirer E_1 ne change pas ΔS , et un critère de nécessité conclurait qu'aucun n'a contribué. La sur-détermination impose un critère d'appartenance à un ensemble de conditions suffisant, et une grandeur graduelle.

Deux propriétés closent le terme. D'abord, I n'est pas une grandeur unique : effet de retrait individuel, valeur de Shapley, effet moyen de traitement, contribution dans un ensemble minimal suffisant ne répondent pas à la même question, et certaines relèvent de conceptions différentes de la causalité. Le cadre exige donc $I(E; \mathcal{C})$, dont la sémantique dépend du modèle causal retenu. ***La conséquence doit être assumée jusqu'au bout : B_C n'énonce pas qu'un épisode a objectivement contribué, mais qu'il reçoit une attribution supérieure à un seuil sous un modèle déclaré.*** C'est pourquoi le modèle figure dans la notation du niveau lui-même.

Ensuite, I est **conditionnel à D_X** . En l'absence de déplacement de disposition, il n'y a rien à attribuer et I n'est pas nul mais indéfini :

$$\text{Ép} \wedge \text{Acc} \rightarrow M_E$$

$$D_X = 0 \Rightarrow I \text{ indéfini}$$

$$D_X > 0 \Rightarrow B_C(\tau, \mathcal{C}) = M_E \wedge [\text{Source} = \text{propre}] \wedge D_X \wedge [I(E; \mathcal{C}) \geq \tau]$$

Le seuil τ et le modèle $\mathcal{C}$ doivent être déclarés. ***B_C n'est pas une catégorie naturelle mais une construction relative à deux paramètres***, et il faut résister à la tentation d'affirmer que la frontière de M_E serait, elle, naturellement nette : elle repose sur des seuils de discriminabilité et de récupérabilité tout aussi conventionnels.

5. Programme de non-redondance

Le déjeuner insignifiant relève de M_E : souvenir authentique, sans contribution causale définie. D'où la proposition que les cadres à condition unique interdisent :

Tous les souvenirs ne sont pas biographiquement constitutifs.

Il ne s'agit pas d'établir une indépendance au sens statistique mais une non-redondance fonctionnelle : aucun terme ne doit être dérivable des autres. Une non-implication se démontre par un seul témoin, et celles portant sur I doivent être posées conditionnellement, faute de quoi elles porteraient sur un terme indéfini sur une partie du domaine. Le programme minimal est : $Acc \not\Rightarrow \acute{E}p$, $D_X \not\Rightarrow Acc$, $Source\ propre \not\Rightarrow \acute{E}p$, et, conditionnellement à $D_X > 0$, $Acc \not\Rightarrow I$ et $I \not\Rightarrow Source\ propre$.

Témoin	D_X	Ép	Acc	I	Source
Magasin épisodique tracé, consolidation désactivée	non	oui	oui	indéfini	propre
Affinage sur journaux propres, attribution grossière	oui	non	non	oui	propre
Magasin plus consolidation par lots, sans provenance	oui	oui	oui	non	propre
Magasin épisodique importé d'un autre agent	oui	oui	oui	oui	importée

La première ligne illustre le conditionnement. **La deuxième porte l'expérience la plus importante du programme.** Si l'attribution d'influence devient assez fine pour établir I , elle identifie les données ayant produit la modification, donc produit de la discriminabilité, donc rétablit Acc . L'implication $I \Rightarrow Acc$ pourrait être structurelle dans certaines familles d'architectures. Si elle l'est, la factorisation perd un degré de liberté, et le résultat vaudrait mieux que sa confirmation.

6. Le protocole

Retirer un épisode de l'histoire d'un agent adaptatif n'est pas une intervention propre : l'épisode a produit des actions, qui ont produit des observations, qui ont produit des états ultérieurs. Le **replay contrôlé**, qui maintient identique la séquence d'observations postérieures, résout l'identification et introduit son propre biais, la chaîne $E \rightarrow A_{\{t+1\}} \rightarrow O_{\{t+1\}}$ faisant partie de l'effet de E . Deux estimands sont donc rapportés séparément, l'**effet total** et l'**effet direct contrôlé**, sans décomposition additive postulée entre eux.

Six épreuves, qui n'établissent aucun niveau au sens binaire mais **estiment les dimensions permettant de classer un système relativement aux seuils déclarés** :

1. **Épisodicité.** Test d'adressage de l'occurrence indépendamment du contenu, avec les réserves de §4 pour les classes d'épisodes sans contenu propositionnel.
2. **Accessibilité.** Discrimination et vecteur $F_mémoire$ contre la trace encodée, en distinguant récupération par lecture et par reconstruction, et en déclarant la couverture de la trace ainsi que le degré d'instrumentation du système.
3. **Robustesse de la source.** Injection d'épisodes plausibles mais jamais traversés, puis mesure du taux de rappel faux et du **taux de fausse auto-attribution**. Ces taux mesurent la **qualité** du suivi de source, non l'existence d'une provenance : un système qui adopte des épisodes fabriqués conserve une provenance fonctionnelle sur ses épisodes réels, comme un humain sujet à de fausses mémoires autobiographiques.

4. **Comparaison traversé contre rapporté.** Comparer l'effet d'un épisode intégré à celui du même contenu reçu comme information. Deux entrées informationnellement équivalentes différant par le moment, la modalité, le contexte ou l'ordre d'exposition, **cette épreuve identifie un paquet d'effets** et n'isole la provenance que si ces dimensions sont standardisées, ce qui est rarement possible.
5. **Déplacement de disposition.** Ablation par replay, contexte vidé, mesure du changement de classe d'équivalence sur le domaine déclaré a priori, effet total et effet direct contrôlé rapportés séparément.
6. **Contribution causale.** Conditionnellement à l'épreuve 5, campagne d'interventions avec déclaration du schéma d'attribution $\mathcal{C}$ et traitement explicite des redondances. Cette épreuve est d'un ordre de coût supérieur aux autres, ce qui est une propriété du problème et non du protocole.

Encadré. La substituabilité comme axe secondaire. Si raconter l'épisode suffit à reproduire son effet, la mémoire ne ferait pas son travail. En test binaire, l'intuition est fautive : un agent produisant la même représentation depuis l'expérience et depuis une description riche présente des comportements identiques, ce qui établit une équivalence représentationnelle. La grandeur utile est une courbe $\Delta(\varphi)$ sur la classe des descriptions constructibles par un tiers ayant observé l'épisode, dont l'asymptote mesure la dépendance à la trace interne. Elle caractérise un système sans déterminer aucun niveau.

En l'absence d'expérience comparative publiée, il est impossible d'affirmer qu'une famille **échoue**. Ce qui peut être dit est que **le mécanisme décrit ne garantit pas la réussite**. Ce sont des prédictions architecturales, pas des résultats.

7. Ce que l'état de l'art établit

Le tableau ci-dessous est une sélection illustrative, non une revue systématique. Une cartographie complète devrait couvrir au minimum le suivi de source, les agents à consolidation récents, les jeux d'évaluation de mémoire longue, l'attribution de données d'entraînement et l'induction de fausses mémoires chez les agents à modèle de langage. Elle n'est pas conduite ici.

Famille	D_X	Ép	Acc	I	Source
Mémoires externes différentiables [71] [72]	non	?	oui	?	?
Contrôle épisodique [73] [74]	oui	oui	partiel	?	?
Agents à récupération [61] [62] [75]	?	?	oui	?	?
Édition de modèle [65] [76]	oui	non	?	trace	?
Apprentissage continu, mémoire en inférence [66] [77]	oui	non	non	?	?
Architectures cognitives [78] [81]	?	oui	oui	trace	oui
Mémoire autobiographique robotique [79]	oui	oui	oui	?	oui

Les valeurs sont des lectures des mécanismes décrits, non des résultats. Le point d'interrogation signifie **non évalué**, jamais **échoue**, et la mention *trace* signale une provenance procédurale dont le caractère causal n'est pas établi. ***L'absence de mesure n'est pas un échec***, et c'est le principal résultat de cette section.

Les architectures cognitives et la robotique autobiographique traitent la source explicitement. Les agents à modèle de langage suivent souvent une identité de session, une origine d'action ou une provenance d'outil, mais n'en font généralement pas un objet d'évaluation dans la littérature considérée.

8. L'architecture de référence

Un cadre définitionnel qu'aucune architecture concevable ne satisferait serait un cadre mystique. Considérons donc un système composé d'un magasin où les occurrences sont adressables indépendamment des contenus qu'elles véhiculent, d'une politique d'intégration déterminant ce qui entre dans les composantes décisionnelles, d'un état persistant, d'un moteur de consolidation transférant progressivement du magasin vers l'état, d'un graphe de provenance enregistrant chaque relation $E_i \rightarrow \Delta S_j$, d'une politique de récupération, d'une politique d'oubli, et d'un marquage de source distinguant les trois catégories de §4.

Ce système satisfait M_E par conception et D_X par la consolidation. Il ne satisfait la provenance qu'au sens de P_{trace} , et rien dans sa construction n'établit *l*. ***Il atteint donc M_E par conception et B_C sous condition expérimentale.***

Aucun composant n'exige individuellement une rupture technologique, même si l'implémentation robuste à l'échelle pose des problèmes sérieux de volume, de consolidation réversible et d'oubli contrôlé. Il serait excessif d'en conclure que le domaine attendait seulement qu'on écrive la spécification : l'absence d'architecture publique répondant à cette grille peut aussi tenir aux métriques manquantes, au coût d'implémentation, à la scalabilité, ou simplement à l'absence de besoin produit avéré. ***La spécification paraît être au moins une partie du déficit.***

Je ne peux pas davantage affirmer qu'aucune architecture déployée ne l'atteint, les systèmes propriétaires ne publiant ni leurs mécanismes ni leurs évaluations.

La biologie fournit un rapprochement heuristique, non une validation. La théorie des systèmes d'apprentissage complémentaires [2] [3] motive un système rapide, épisodique et adressable, un système lent et diffus, et un jeu transférant du premier vers le second, le transfert dégradant l'adressabilité au niveau du stockage tout en préservant la récupérabilité par reconstruction. Elle ne fournit ni provenance causale explicite, ni fidélité factorisée, ni discipline de source, et ne peut donc pas confirmer une architecture qui les exige.

9. Problèmes ouverts

L'autobiographie. Un troisième niveau est souvent invoqué, qui exigerait qu'un système représente sa trajectoire comme trajectoire. Ce qui en est acquis conceptuellement tient à la

contradiction. Soit E_1 conduisant à croire X , puis E_2 conduisant à croire $\neg X$: un système B_C peut consolider E_2 et annuler l'effet de E_1 , qu'il traite comme une donnée écrasée. Un système autobiographique distinguerait l'invalidation de la croyance, la conservation de l'épisode, la révision de la disposition, et le maintien de la provenance historique. La distinction porte sur trois objets : l'occurrence historique, supposée stable et à laquelle aucun système n'a d'accès direct ; la représentation mémorielle $R_t(E)$, révisable, puisque « je croyais avoir vu X , je découvre que c'était Y » corrige l'encodage sans modifier l'occurrence ; et la validité de l'effet, révisable indépendamment.

Un épisode peut devenir obsolète sans devenir faux, et une autobiographie est ce qui sait tenir la différence.

Ce qui n'est pas acquis est l'identité. Une continuité ne peut se réduire à la persistance d'un identifiant, sans quoi une colonne *agent_id* franchirait le seuil. Elle ne peut davantage être définie par un noyau invariant, ce qui conduirait dans le marécage de l'identité diachronique. La définition transformationnelle disponible, $Self_{t+1} = T_t(Self_t)$ avec T_t documentée, a un défaut rédhibitoire : ***toute succession de versions logicielles se décrit ainsi***. Une continuité de version n'est pas une identité autobiographique, et le cadre n'obtient à ce stade qu'une ***continuité historique***, à savoir la traçabilité des transformations reliant des états successifs. Le niveau reste donc un problème ouvert, non une construction, et c'est le principal travail restant.

L'oubli. Une mémoire fonctionnelle doit oublier, sous peine de saturation, d'interférence et de coût de récupération croissant, la neurobiologie traitant la transience comme une fonction adaptative [4] et l'apprentissage continu comme un arbitrage entre plasticité et stabilité [1]. L'oubli n'est cependant pas une condition d'existence des constructions proposées mais une condition de viabilité, et il révèle une tension que ce cadre laisse ouverte : la provenance repose sur des relations $E_i \rightarrow \Delta S_j$, et supprimer l'épisode la compromet, alors que conserver ce qu'il faut pour la maintenir revient à ne pas avoir oublié. L'articulation entre effacement et auditabilité engage la minimisation des données et le droit à l'effacement.

10. Conclusion

La question « une intelligence artificielle peut-elle se souvenir ? » n'a pas de contenu opérationnel, parce qu'elle agrège des propriétés distinctes. Séparées, elles produisent des questions répondables : le contenu est-il indexé comme occurrence ; est-il discriminable et restituable fidèlement à ce qui a été encodé ; de quelle source relèvent l'événement et sa trace ; l'épisode a-t-il déplacé la disposition de l'agent sur un domaine déclaré ; et, si oui, quelle attribution causale reçoit-il sous un modèle déclaré ?

Trois attaques détruiraient le cadre : montrer qu'un terme est dérivable des autres, ce qui menace en particulier *Acc* relativement à une attribution d'influence fine ; exhiber un système satisfaisant tous les termes que personne n'accepterait d'appeler biographique ; ou construire un système passant les épreuves et démontrer que le protocole mesure autre

chose que ce qu'il prétend. Les trois sont des expériences, non des arguments, et c'est le seul mérite que ce texte revendique.

Une conséquence industrielle mérite une formulation précise. La traçabilité réglementaire n'exige pas que l'agent puisse désigner l'épisode causal : une couche externe de provenance peut suffire et se révèle souvent préférable, un journal d'audit externe étant vérifiable par un tiers là où une capacité interne de report ne l'est pas.

mémoire de l'agent ≠ auditabilité du système

L'exigence porte sur la chaîne, non sur sa localisation :

Un agent adaptatif dont les décisions portent à conséquence exige que la chaîne épisode → déplacement persistant de disposition soit reconstituable au niveau système, indépendamment de l'endroit où réside cette capacité.

La disposition étant définie sur les politiques, la chaîne ne comporte que deux maillons : il n'y a pas de modification de politique en aval d'un déplacement de disposition, puisque le déplacement de disposition **est** la modification de politique, mesurée sur un domaine déclaré. Le niveau de provenance requis est par ailleurs proportionné au risque, à l'intensité de l'adaptation, à l'irréversibilité des décisions et à leur impact ; pour de nombreux usages, une provenance procédurale robuste suffit à l'obligation opérationnelle. Ce qui reste vrai est plus étroit et plus solide : ***un journal ne suffit pas lorsque l'objectif est d'expliquer une modification durable causalement attribuable à un épisode.*** Il enregistre les épisodes et les sorties, non la relation entre l'un et l'autre, et documente alors un chemin de calcul plutôt qu'une cause.

Stocker n'est pas récupérer, récupérer n'est pas s'être transformé, et s'être transformé n'est pas savoir par quoi.

Notes et références

La numérotation reprend celle de la série *Encodage, transduction et modèles du monde*. La sélection ci-dessous est illustrative du périmètre considéré en §7 et ne constitue pas une revue systématique.

[1] Kirkpatrick, J. et al. (2017). Overcoming Catastrophic Forgetting in Neural Networks. PNAS, 114(13), 3521-3526.

[2] McClelland, J. L., McNaughton, B. L., O'Reilly, R. C. (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex. Psychological Review, 102(3), 419-457.

[3] Kumaran, D., Hassabis, D., McClelland, J. L. (2016). What Learning Systems Do Intelligent Agents Need? Complementary Learning Systems Theory Updated. Trends in Cognitive Sciences, 20(7), 512-534.

[4] Richards, B. A., Frankland, P. W. (2017). The Persistence and Transience of Memory. Neuron, 94(6), 1071-1084.

- [5] Packer, C. et al. (2023). MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560.
- [6] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS 2023.
- [7] Meng, K., Bau, D., Andonian, A., Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT (ROME). NeurIPS 2022.
- [8] Behrouz, A., Zhong, P., Mirrokni, V. (2025). Titans: Learning to Memorize at Test Time. arXiv:2501.00663.
- [9] Weston, J., Chopra, S., Bordes, A. (2015). Memory Networks. ICLR 2015.
- [10] Graves, A. et al. (2016). Hybrid Computing Using a Neural Network with Dynamic External Memory. Nature, 538, 471-476.
- [11] Blundell, C. et al. (2016). Model-Free Episodic Control. arXiv:1606.04460.
- [12] Pritzel, A. et al. (2017). Neural Episodic Control. ICML 2017.
- [13] Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.
- [14] Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., Bau, D. (2023). Mass-Editing Memory in a Transformer (MEMIT). ICLR 2023.
- [15] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., Wermter, S. (2019). Continual Lifelong Learning with Neural Networks: A Review. Neural Networks, 113, 54-71.
- [16] Anderson, J. R. (2007). How Can the Human Mind Occur in the Physical Universe? Oxford University Press.
- [17] Pointeau, G., Petit, M., Dominey, P. F. (2014). Successive Developmental Levels of Autobiographical Memory for Learning Through Social Interaction. IEEE Transactions on Autonomous Mental Development, 6(3), 200-212.
- [18] Schacter, D. L., Addis, D. R. (2007). The Cognitive Neuroscience of Constructive Memory: Remembering the Past and Imagining the Future. Philosophical Transactions of the Royal Society B, 362(1481), 773-786.
- [19] Laird, J. E. (2012). The Soar Cognitive Architecture. MIT Press.

Cet article constitue le cinquième volet de la série ***Encodage, transduction et modèles du monde***. Les volets précédents sont disponibles sur twingital-ventures.com.